UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

COMPREENSÃO DE DIFERENÇAS CULTURAIS PARA IDENTIFICAR PESSOAS COM POTENCIAIS INTERESSES COMUNS EM SITES DE REDES SOCIAIS

GILBERTO ASTOLFI

SÃO CARLOS
2010

COMPREENSÃO DE DIFERENÇAS CULTURAIS PARA IDENTIFICAR PESSOAS COM POTENCIAIS INTERESSES COMUNS EM SITES DE REDES SOCIAIS

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GILBERTO ASTOLFI

COMPREENSÃO DE DIFERENÇAS CULTURAIS PARA IDENTIFICAR PESSOAS COM POTENCIAIS INTERESSES COMUNS EM SITES DE REDES SOCIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de mestre em Ciência da Computação.
Orientação: Profa. Dra. Junia Coutinho Anacleto

SÃO CARLOS
2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

A856cd

Astolfi, Gilberto.
Compreensão de diferenças culturais para identificar pessoas com potenciais interesses comuns em sites de redes sociais / Gilberto Astolfi. -- São Carlos : UFSCar, 2011.
126 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Ciência da computação. 2. Senso comum. 3. Análise semântica. 4. Cultura. 5. Cultura popular. I. Título.

CDD: 004 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

"Compreensão de diferenças culturais para identificar pessoas com potenciais interesses comuns em Sites de Redes Sociais"

GILBERTO ASTOLFI

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Profa. Dra. Júnia Coutinho Anacleto
(Orientadora - DC/UFSCar)

Profa. Dra. Vânia Paula de Almeida Neris
(DC/UFSCar)

Prof. Dr. Sérgio Roberto Pereira da Silva
(DIN/UEM)

São Carlos
Setembro/2010

Dedico esta Dissertação à Deus,
à minha família, em especial meu pai (em memória)
e minha mãe, que juntos me educaram
para que eu chegasse até aqui.

AGRADECIMENTOS

Agradeço primeiramente à Deus por ter me dado muita sorte em tudo que faço.

Obrigado minha mãe e meu pai (em memória) pela educação, serenidade e muitos outros ensinamentos que me passaram para que eu pudesse alcançar meus objetivos, garanto que sem isso não estaria chegado até aqui.

Agradeço meus irmãos pela dedicação e esforço e muitas outras coisas que fizeram por mim ao longo de toda a minha vida.

À minha esposa pela paciência, dedicação e por se privar de muitas coisas para que eu conseguisse atingir essa meta em minha vida.

Agradeço à Júnia pela grande oportunidade de mudar minha história de vida e pelos ensinamentos ao longo de todo o período em que convivemos. Pode ter certeza que nunca me esquecerei de sua imensa ajuda.

Agradeço a todos os colegas de laboratório, Ana, Johana, David, Bruno e Fernando pelas dicas e auxílio nas decisões durante a realização deste trabalho. Quero agradecer também a Vanessa pela grande ajuda nas publicações de artigos, sem ela os resultados deste trabalho não seriam publicados. Ao Marcos, um grande homem e amigo, que me ajudou muito, muito mesmo, em todas as fases deste trabalho, sem a ajuda dele, certamente, este trabalho não chegaria ao grau de maturidade que está hoje.

Quero agradecer muito meus colegas de trabalho, que posso chamá-los de amigos, João Marcos, Jossimar e Angélica. Eles, incondicionalmente, não mediram esforços para me dar o suporte necessário para eu me dedicar somente ao curso.

Enfim, agradeço a todos que de alguma forma contribuíram com este trabalho e minha história, cada um foi muito importante para minha vida. Meu muito Obrigado!

RESUMO

Este trabalho descreve a obtenção de um método, por meio de uma metodologia cíclica de trabalho, que objetiva identificar pessoas, usuários de redes sociais, que estão falando sobre um mesmo assunto. A motivação para a obtenção deste método partiu de um problema percebido em Sistemas Gerenciadores de Redes Sociais, que é identificar e/ou agrupar pessoas em torno de um mesmo assunto, a fim poder viabilizar uma possível ligação social entre elas. O diferencial desse método está relacionado em considerar a cultura das pessoas como o ponto principal para tentar identificar as diversas formas que elas têm para se expressar sobre um determinado assunto e, assim, observar quais são as pessoas que estão escrevendo sobre as mesmas coisas nas redes sociais, considerando as muitas formas como elas se expressam. Nesse contexto, esse método faz uso das informações culturais obtidas por meio do Projeto Open Mind Common Sense no Brasil (OMCS-Br), que tem como objetivo coletar o conhecimento cultural dos brasileiros (ANACLETO, 2008), a fim de identificar similaridade de contextos entre vocabulários distintos. Com o intuito de observar a viabilidade desse método, foi realizado um estudo de caso com a participação de pessoas, que por meio de questionários puderam opinar sobre os resultados obtidos pelo método, com o intuito de observar a viabilidade de sua aplicabilidade.

Palavras-chave: Senso Comum. Redes Social. Comparações Semânticas Textuais. Contexto. Cultura.

ABSTRACT

In this research we described a way to obtain a method, through a cyclical approach of work, which goals to identify people, users of social networks that are talking about the same subject. The motivation for obtaining this method came from the perceived problem in the Social Network Management Systems, which is to identify and/or group people around the same issue in order to make social touch possible among them. The differential this method is related to consider people's culture as the main point to try to identify the various forms they have to express themselves on a particular issue, and thereby to observe what people are writing about it in the social networks, considering the many ways in which people express themselves. In this context, this method makes use of cultural information get by the Open Mind Common Sense in Brazil Project (OMCS-Br), that goal to collect Brazilians' cultural knowledge (ANACLETO, 2008), in order to identify similarity of contexts among different vocabularies. With the purpose of to observe the feasibility of this method, we performed a case study with the participation of people that through questionnaires could opine on the results obtained by the method, in order to observe the feasibility of its applicability.

Palavras-chave: Common Sense. Social Networking. Semantic Textual Comparisons. Context. Culture.

LISTA DE FIGURAS

Figura 1. 1 Metodologia de trabalho adotada. 3
Figura 2. 1 Histórico de lançamento dos principais SNSs Boyd (2007). 8
Figura 2. 2 Página do perfil de um usuário do SNS Migente. 9
Figura 2. 3. Página do perfil de um usuário do SNS Orkut. 10
Figura 2. 4.Sugestões de amizade do Orkut. 13
Figura 3. 1. Arquitetura do Projeto OMCS-Br. 20
Figura 3. 2. Temas e atividades que podem ser usadas pelo colaborador do Projeto OMCS-Br. 21
Figura 3. 3. Template para coleta de conhecimento cultural. 21
Figura 3. 4. Rede semântica – ConceptNet. 23
Figura 3. 5. Fase de extração. 24
Figura 3. 6. Fase de Normalização. 25
Figura 3. 7. Fase de Relaxamento. 25
Figura 3. 8. Exemplo de associação entre a relação semântica e o perfil do colaborador. 26
Figura 3. 9. Exemplo de uso da API da ConceptNet. Esse aplicativo foi desenvolvido por David Buzatto. 27
Figura 4. 1. (a) parte dos conceitos obtidos por meio da busca na base do OMCS-Br com o perfil dos colaboradores. (b) estado da rede semântica depois de extraído os conceitos com maior frequência. 32
Figura 4. 2. Estado da rede semântica depois dos conceitos com maior frequência serem retirados. 33
Figura 4. 3. Simulação da busca do perfil que mais representa certo vocabulário. Conceitos com maior frequência retirados da rede semântica. 34
Figura 4. 4. Arquitetura do método quando considera o perfil e o vocabulário das pessoas. ..36
Figura 4. 5. Desempenho entre os assuntos usados no experimento a fim de identificar pessoas que possuem o mesmo perfil e compartilham o mesmo vocabulário. 39
Figura 4. 6. Comparação entre os desempenhos dos assuntos considerando apenas o perfil dos usuários do 39
Figura 4. 7. Exemplo de postagem nas comunidades do Orkut. 42
Figura 4. 8. Exemplo de uma associação implícita entre conceitos. 46
Figura 4. 9. Coeficiente de Jaccard. 49
Figura 4. 10. Exemplo de uma postagem recuperada por meio de buscas por pares de palavras. 50
Figura 4. 11. Arquitetura da busca do método, quando considera apenas o vocabulário das pessoas. 50
Figura 4. 12. Representação gráfica de *mr*. 57
Figura 4. 13. Exemplo do retorno de uma análise feita pelo PALAVRAS. 59
Figura 4. 14. Exemplo gráfico de uma análise sintática feita pelo PALAVRAS. 60
Figura 4. 15. Exemplo da atuação do algoritmo sobre uma análise feita pelo PALAVRAS. .61
Figura 4. 16. Diagrama de classes representando uma *mr*. 64
Figura 4. 17. Exemplo gráfico da representação de uma *mr*. 64
Figura 4. 18. Representação de conhecimento composta por um conjunto de mr geradas a partir de um conjunto de orações. 65
Figura 4. 19. Representação de conhecimento enriquecida com conhecimento cultural. 66
Figura 4. 20. Representação de conhecimento após o uso de banco de sinônimos. 66
Figura 4. 21. . Relações semânticas geradas a partir de um conjunto de orações que expressam um assunto. 67

Figura 4. 22. Parte de representação de conhecimento sobre um assunto. 68
Figura 4. 23. Exemplo de uma busca usando uma mr pertencente a representação de conhecimento. 68
Figura 4. 24. Exemplo de recuperação do trecho de texto de uma postagem que possivelmente, depois de transformado em mr, pode ser igual a mr em questão usada na busca. 68
Figura 4. 25. Relação entre o desempenho das mr geradas a partir das orações com o apoio da base do OMCS e do apoio da base de sinônimos. 72
Figura 4. 26. Representação gráfica da relação IsA. 77
Figura 4. 27. Representação gráfica da relação DefinedAs. 78
Figura 4. 28. Dois tipos possíveis de busca de conhecimento cultural na base do OMCS-Br. 79
Figura 4. 29. Submissão dos componentes de mr como referência ao OMCS-Br, a fim de identificar sinônimos culturais. 80
Figura 4. 30. Exemplo de concepts depois de finalizar a expansão de uma mr com conhecimento cultural. 82
Figura 4. 31. Arquitetura do método que identifica pessoas que falam sobre o mesmo assunto em SNSs. 83
Figura 4. 32. Comparação do desempenho entre as mr geradas a partir da expansão semântica e a mr gerada a partir da oração. 85
Figura 5. 1. Exemplos de perguntas do questionário sobre o perfil do participante 89
Figura 5. 2. Perguntas do questionário para identificar a familiaridade dos participantes com a leitura. 90
Figura 5. 3. Exemplo do questionário usado para observar o uso método. 91
Figura 5. 4. Exemplo de como o questionário foi aplicado ao participante do estudo de caso. 95
Figura 5. 5. Comparação entre a quantidade de postagens recuperadas com e sem o apoio de conhecimento cultural. 97
Figura 5. 6. Resultado referente às repostas da Pergunta 1 para os três assuntos usados no estudo de caso. 99
Figura 5. 7. Comparação entre o resultado do “uso” e o “não uso” de conhecimento cultural em relação às respostas da Pergunta 1. 101
Figura 5. 8. Resultado referente às repostas da Pergunta 2 para os três assuntos usados no estudo de caso. 104
Figura 5. 9. Comparação entre o resultado do “uso” e o “não uso” de conhecimento cultural em relação às respostas da Pergunta 2. 104
Figura 5. 10. Resultado referente às repostas da Pergunta 3 para os três assuntos usados no estudo de caso. 106
Figura 5. 11. Comparação entre o resultado do “uso” e o “não uso” de conhecimento cultural em relação às respostas da Pergunta 3. 107
Figura 5. 12. Resultado referente às repostas da Pergunta 4 para os três assuntos usados no estudo de caso. 109
Figura 5. 13. Comparação entre o resultado da Pergunta 1, Pergunta 2 e Pergunta 3. 110

LISTA DE TABELAS

Tabela 2. 1 Relação dos principais SNSs ordenados pela quantidade de usuários. 11
Tabela 3. 1. Relações semânticas que compõem o projeto OMCS-Br (LIU, 2004). 23
Tabela 4. 1. Exemplo de perfil extraído de certo SNS. 34
Tabela 4. 2. Resultado da busca por PC para cada assunto. 37
Tabela 4. 3. Definição das relações de Minsky da classe das K-lines. 45
Tabela 4. 4. Exemplo de PC. Uma reprodução de uma linha da Tabela 4.2. 46
Tabela 4. 5. Exemplo de uma busca por conceitos na base do OMCS-Br usando somente a classe das K-lines. 48
Tabela 4. 6. Quantidade de postagens identificadas e o número de palavras de concepts existente em cada uma delas. 52
Tabela 4. 7. Exemplos de *mr* geradas a partir de orações. 58
Tabela 4. 8. Listagem das etiquetas usadas pelo PALAVRAS consideradas por este trabalho. 60
Tabela 4. 9. Associação dos componentes de uma mr com as etiquetas usadas pelo PALAVRAS. 63
Tabela 4. 10. mr oriundas das orações usadas como sementes. 70
Tabela 4. 11. mr geradas a partir do auxílio da base de conhecimento cultural do projeto OMCS-Br. 70
Tabela 4. 12. mr geradas a partir do uso da base de sinônimos. 70
Tabela 4. 13. Listagem do desempenho de cada mr geradas a partir do uso da oração, base cultural do OMCS-Br e da base de sinônimos. 71
Tabela 4. 14. Distribuição da faixa etária dos colaboradores do projeto OMCS-Br. 75
Tabela 4. 15. Distribuição dos colaboradores do projeto OMCS-Br por região. 75
Tabela 4. 16. Distribuição dos colaboradores do projeto OMCS-Br por nível de escolaridade. 76
Tabela 4. 17. Exemplo de expansão semântica de metas-relação. 78
Tabela 4. 18. Resultado da busca na base de conhecimento cultural do OMCS-Br usando as relações IsA e DefinedAs. 81
Tabela 4. 19. Exemplo de mr geradas a partir de uma oração, considerando a expansão semântica. 84
Tabela 4. 20. Ranking entre as mr com melhores resultados durante as buscas nas comunidades do Orkut. 85

SUMÁRIO

1 INTRODUÇÃO 1

1.1 Motivação 1
1.2 Problema 1
1.3 Objetivo 2
1.4 Abordagem 3
1.4.1 Abordagem para obtenção do método 3
1.4.2 Abordagem para a validação do método 4
1.5 Organização do trabalho 5

2 SITES DE REDES SOCIAIS 6

2.1 Considerações iniciais 6
2.2 Os SNSs 6
2.3 A Busca por pessoas nos SNSs 12
2.4 Trabalhos relacionados com a identificação de pessoas que possam promover à aproximação social 14
2.5 Considerações finais 17

3 O PROJETO OPEN MIND COMMON SENSE NO BRASIL 19

3.1 Considerações iniciais 19
3.2 O Projeto OMCS-Br 19
3.3 O Site 20
3.4 A geração da ConceptNet 24
3.5 O perfil dos colaboradores 26
3.6 Acesso a ConceptNet 26
3.7 Considerações finais 27

4 ETAPAS PARA A OBTENÇÃO DO MÉTODO 28

4.1 Considerações iniciais 28
4.2 Iteração 1 – Identificando pessoas com o mesmo perfil e vocabulário 29
4.2.1 Exposição do problema 29
4.2.2 Resolução do problema 30
4.2.3 Teste 36
4.2.4 Apontar falhas 40
4.3 Iteração 2 – Identificando pessoas com um vocabulário comum 41

4.3.1 Comunidades do Orkut 42
4.3.2 Exposição do problema 43
4.3.3 Resolução do problema 44
4.3.4 Teste 51
4.3.5 Apontar falhas 53
4.4 Iteração 3 – Identificando pessoas que falam sobre o mesmo assunto 54
4.4.1 Exposição do problema 54
4.4.2 Resolução do problema 55
4.4.3 Teste 69
4.4.4 Apontar falhas 72
4.5 Iteração 4 – Identificando pessoas que falam sobre o mesmo assunto tornando as diferenças culturais irrelevantes. 73
4.5.1 Exposição do problema 73
4.5.2. Resolução do problema 74
4.5.3 Teste 84
4.5.4 Apontar falhas 85
4.6 Considerações finais 86

5 ESTUDO DE CASO 87

5.1 Considerações iniciais 87
5.2 Planejamento do estudo de caso 87
5.3 Etapas do estudo de caso 89
5.4 Questionários utilizados no estudo de caso 89
5.4.1 Questionário usado para coletar o perfil dos participantes 89
5.5 Base para a elaboração dos questionários 91
5.6 Primeira etapa do estudo de caso 92
5.6.1 Resultados dessa etapa 92
5.7 Segunda etapa do estudo de caso 95
5.8 Terceira etapa do estudo de caso 96
5.8.1 Análise sobre a Pergunta 1 98
5.8.2 Análise sobre a Pergunta 2 102
5.8.3 Análise sobre a Pergunta 3 105
5.8.4 Análise sobre a Pergunta 4 107
5.9 Considerações finais 111

6 CONCLUSÕES E TRABALHOS FUTUROS 112

6.1 Síntese dos principais resultados.....112
6.2 Trabalhos futuros.....113

7 REFERÊNCIAS 115

APÊNDICES

APÊNDICE A - UM ESTUDO SOBRE AS RELAÇÕES DE MINSKY.....120
APÊNDICE B - QUESTIONÁRIO PARA COLETA DE PERFIL.....125
APÊNDICE C - QUESTIONÁRIO PARA AVALIAR O MÉTODO PROPOSTO.....126

1 INTRODUÇÃO

1.1 Motivação

Atualmente, os sites de redes sociais (em inglês, Social Networking Systems - SNSs) têm se tornado uma ferramenta utilizada por muitas pessoas. Elas utilizam essa ferramenta para encontrar e conversar com os amigos, estudar, discutir um determinado assunto, jogar e tantas outras atividades relacionadas com o entretenimento, estudo, trabalho, etc.

A quantidade de usuários, que ultrapassa centenas de milhões (WEAVER, 2008), e todos os recursos que fazem ou podem fazer parte dos SNSs são alguns dos fatores que têm despertado e/ou aumentado o interesse da comunidade cientifica e do mercado, pois ambos querem investigar o que acontece nessas redes sociais para que as pessoas se interessem e as utilizem tanto, bem como perceber novos recursos ou formas de aprimorá-los com o intuito de facilitar o uso ou proporcionar aos usuários novas formas de interagir com os mesmos, podendo assim formar redes sociais mais ativas, etc., sendo assim, há o interesse de estudar os fenômenos inerentes a esse tema (SOCIALNET, 2009; WAIHCWS, 2010).

1.2 Problema

Um dos recursos principais dos SNSs é permitir que as pessoas se encontrem, no entanto, essa tarefa nem sempre é atingida de modo satisfatório. Apesar do campo de busca, em que o usuário pode tentar localizar outro usuário, um conteúdo, etc., ser algo comum entre os SNSs nem sempre é possível encontrar o que se quer.

Um potencial problema para essa falha na localização de pessoas, etc., está na forma com que a busca é realizada, pois os mecanismos de buscas, na maioria das vezes, comparam o que foi digitado pelo usuário, no campo de busca, com o que existe armazenado no SNS. Por exemplo, se o usuário digita “Lula”, ele pode ter como resultado todas as pessoas que falam sobre esse assunto, no entanto, ele tem acesso apenas as pessoas que falam desse assunto com essa palavra “Lula”, sendo assim, o usuário não terá acesso às pessoas que falam sobre “presidente”, “presidente do Brasil” e “governo petista”. Essa limitação faz com que os usuários não tenham acesso a todas as pessoas que se interessam pelo mesmo assunto, pois a forma com que cada um escreve influencia diretamente no resultado que obterá da busca.

A forma com que cada pessoa escreve e se expressa é influenciada pelo o que ela teve a oportunidade de aprender e conhecer, pela convivência com as demais pessoas,

enfim, pela cultura, por isso, uma possibilidade de melhorar os mecanismos de busca, é considerar a cultura das pessoas nesse processo. A cultura deve ser considerada como um potencial para unir as pessoas com diferentes experiências, conhecimentos, etc., e não para distanciá-las.

1.3 Objetivo

O objetivo deste trabalho foi desenvolver um método para identificar pessoas em SNSs que estão falando sobre o mesmo assunto, por meio de comparações semânticas textuais, considerando a cultura das mesmas. Cultura que, Segundo LARAIA (2006), é um complexo que inclui o conhecimento, as crenças, a arte, a moral, a lei, língua falada e escrita, os costumes e todos os outros hábitos e aptidões adquiridos pelo homem como membro da sociedade.

Com esse método, é possível considerar a cultura das pessoas para identificar o que cada uma está falando nos SNSs, para assim, pesquisar outras pessoas que estão falando sobre o mesmo assunto, mesmo de formas diferentes. Nesse contexto, as pessoas que estão falando sobre um determinado assunto poderiam ter a chance de conhecer umas as outras para conversarem, discutirem, ou seja, aprenderem umas com as outras por meio da troca de experiência, conhecimento, enfim, cultura.

Para considerar a cultura das pessoas o método aqui proposto faz uso de conhecimento cultural, obtido por meio do Projeto Open Mind Common Sense no Brasil (OMCS-Br), que tem como objetivo coletar o conhecimento cultural dos brasileiros (ANACLETO, 2008).

Com o intuito de observar o uso desse método foi realizado um estudo de caso para que pessoas pudessem analisar os resultados obtidos e, assim relatarem se os usuários identificados estão falando sobre o mesmo assunto, mesmo de formas diferentes, bem como avaliar se as comparações semânticas textuais possuem um bom desempenho quando aplicadas nesse contexto.

Além disso, foi avaliado o quanto os usuários identificados poderiam possuir interesse a um determinado assunto em questão e, se há a possibilidade de utilizar o método para apoiar Sistemas de Recomendação Social (TERVEEN, 2005). Ressalta-se que essas duas últimas avaliações não são o enfoque deste trabalho. Os dados coletados são apenas para observar alguns indícios que serão importantes para os trabalhos futuros.

1.4 Abordagem

1.4.1 Abordagem para obtenção do método

Neste trabalho foi adotada a estratégia experimental para a obtenção do método (WAZLAWICK, 2009), pois ao longo de seu desenvolvimento, foram realizadas iterações ou mudanças sistemáticas para verificar o reflexo das alterações, sempre buscando melhores resultados.

Um arcabouço, baseado no modelo Espiral de Engenharia de Software proposto por Boehm (1986), foi modelado e adotado como abordagem de desenvolvimento do trabalho, a fim de apoiar a sistematização exigida pelo modelo de pesquisa experimental.

No modelo Espiral de Engenharia de Software a cada iteração ao redor do espiral, versões progressivamente mais complexas do software são construídas. A abordagem de trabalho adotada aqui segue esse mesmo raciocínio, mas ao invés de versões de produto de software, novas versões do método foram entregues, com o intuito de suprir falhas na versão anterior.

A Figura 1.1 mostra as quatro fases de uma iteração. Observa-se que elas compõem um processo que é baseado em identificação e resolução de problemas.

Figura 1. 1 Abordagem de trabalho adotada.

Expor o problema – Essa fase tem o objetivo de mostrar os problemas que ocorreram na iteração anterior. Caso seja a primeira, ela se baseia no problema de pesquisa identificado.

Resolução do problema – Nessa fase é apresentada qual a estratégia adotada para solucionar o problema identificado na fase anterior. Os métodos, as ferramentas e como proceder são escolhidos de forma livre, isto é, pode ou não sofrer influências da iteração anterior. Isso depende de estudos realizados na busca da solução do problema a ser resolvido, e na regra levantada para se chegar ao objetivo.

Testar a solução ou Experimentação – Nessa fase os esforços são concentrados em testes, que são realizados para verificar se a abordagem adotada para a resolução do problema conseguiu obter avanço. Ela também identifica o que deve ser controlado e observado.

Apontar falhas – Nessa fase são verificados os dados oriundos dos testes para averiguar se houve melhora nos resultados em relação à iteração anterior. Aqui são apontados os problemas e ganhos conseguidos para gerar subsídios para a nova iteração de refinamento do método.

1.4.2 Abordagem para a validação do método

Neste trabalho foi adotada a estratégia de estudo de caso para verificar a viabilidade do método proposto. O estudo de caso, de acordo com Yin (2002) é um tipo de investigação empírica, que sobre um fenômeno inserido em um contexto da vida real, permite um estudo mais aprofundado.

Segundo Dias (2000), o estudo de caso consiste em uma investigação a fim de prover uma análise do contexto e das particularidades envolvidas no fenômeno em estudo. A vantagem é que o fenômeno não está isolado de seu contexto (como nas pesquisas de laboratório), por esse motivo que se adotou neste trabalho esta estratégia para a validação do método.

O estudo de caso pode incluir evidências qualitativas bem como evidências quantitativas. Na análise qualitativa o objetivo é adquirir conhecimento e dar significado a uma determinada experiência por meio da organização, interpretação e categorização dos dados. Já na quantitativa a observação é mais controlada e utiliza métodos quantitativos na análise e descrição do fenômeno (DIAS, 2000).

Neste trabalho adotou-se a análise quantitativa dos dados obtidos através da realização do estudo de caso, pois levando em consideração a proposta, essa abordagem se faz ideal, uma vez que se deseja observar em um contexto da vida real, ou seja, redes sociais.

1.5 Organização do trabalho

Este trabalho encontra-se organizado em seis capítulos.

Capítulo 2 Discute uma análise sobre o crescimento, o uso e as funcionalidades dos principais SNSs da atualidade. Além principais pesquisas relacionadas à identificação de contextos que possam inferir ligações sociais on-line entre pessoas.

Capítulo 3 Apresenta uma síntese literária sobre os conceitos relacionados ao Projeto OMCS-Br (*Open Mind Common Sense* no Brasil), que é utilizado para coletar, armazenar, processar e disponibilizar conhecimento cultural para aplicações computacionais.

Capítulo 4 Descreve todas as iterações que foram instanciadas na abordagem de trabalho para a obtenção do método proposto por este trabalho. Aqui são detalhadas as decisões e estratégias para um melhor refinamento da proposta.

Capítulo 5 Apresenta o estudo de caso realizado para observar o uso do método proposto por este trabalho. Nesse estudo houve a participação de pessoas para avaliar o resultado obtido por meio desse método.

Capítulo 6 Discute as principais informações obtidas por meio do estudo de caso, além de apontar trabalhos futuros.

2 SITES DE REDES SOCIAIS

2.1 Considerações iniciais

Nos últimos anos, milhões de usuários têm sido atraídos para os SNSs, sendo que muitos desses usuários têm esses sistemas como parte de seu cotidiano, usando-os para entretenimento, negócios, estudos, etc. (HOWARD, 2008).

Atualmente existem centenas de SNSs, sendo que alguns são voltados para públicos específicos, como por exemplo, para religião (www.mychurch.com), empresas (Beehive) e grupos étnicos (www.migente.com) e outros para público geral, como o Orkut (www.orkut.com), o Hi5 (www.hi5.com) e o Facebook (www.facebook.com). A maioria desses SNSs disponibiliza diferentes serviços e ferramentas, tais como, conectividade móvel, compartilhamento de vídeos e fotos, blogs, etc., ajudando os usuários a construir e manter suas redes sociais.

Nesse capítulo são discutidos os principais SNSs ativos da atualidade, detalhando sua evolução, foco, popularidade e como alguns deles são usados.

2.2 Os SNSs

Um SNS é constituído de serviços que permitem que usuários construam um perfil, mantendo uma lista de outros usuários com quem compartilham laços de amizade, podendo também visualizar a lista de contatos desses usuários com quem se relaciona (WEAVER, 2008). Os SNSs fornecem aos usuários os meios que permitem a interação entre eles possibilitando que haja uma conexão entre os indivíduos que antes não era possível, seja pela distância geográfica ou pelo idioma. De acordo com Anderson (1992), quando uma pessoa constrói seu perfil, ela primeiramente procura manter pessoas já conhecidas – na "vida real" – em sua lista de contatos.

Apesar dos SNSs possuírem várias ferramentas, a principal delas é a de gerenciamento do perfil do usuário (HOWARD, 2008). Esse gerenciamento se dá por meio da utilização de páginas de conteúdo onde estão todos os dados do usuário, que são preenchidos durante a criação do perfil e que podem ser editados posteriormente. Geralmente o perfil é composto por alguns dados, como nome, localização, idade, sexo, foto, entre outros.

Para que um usuário possa aumentar sua lista de contatos, os SNSs permitem que ele formalize um convite, fazendo uma "solicitação de amizade" ou uma "solicitação de contato", dependendo do SNS, a outro usuário que não faz parte da lista de contatos.

O indivíduo convidado pode aceitar – ou não – o convite. Caso o convite seja aceito, é então criada uma ligação entre eles, que agora vão fazer parte de uma mesma rede social. Não há muitos estudos que definem um padrão na forma que um usuário segue na escolha de uma nova ligação, entretanto alguns estudos empíricos, como de Crandall *et al.* (2008), mostram que um usuário tende a escolher novas ligações a partir de pessoas semelhantes à ele, praticando assim o processo de seleção. Além disso, Crandall et al. (2008) também dizem que a influência social, isto é, pessoas mais influentes em um determinado grupo social, também tem peso na decisão de uma ligação de amizade.

Mislove *et al.* (2008) afirmam que uma pessoa, quando procura novas ligações, tende a escolher usuários com muitos amigos. Dessa forma, usuários com muitas ligações tendem a ter mais, ou seja, as chamadas ligações preferenciais, que são popularmente definidas como: “os ricos ficam mais ricos”.

Em muitos SNSs, como o Orkut e o Hi5, há a possibilidade da formação de grupos de pessoas em torno de um interesse comum. Esses grupos são chamados de comunidades e, para que um indivíduo possa participar de uma, ele deve estar cadastrado no SNS. As comunidades não ficam restritas somente a indivíduos que fazem parte de uma mesma rede social, elas podem ser formadas por indivíduos de redes sociais diferentes e que não mantém nenhum tipo de laço social entre si.

Boyd (2007) apresenta um histórico interessante sobre os lançamentos de SNSs, sendo que o precursor foi o SixDegreess lançado em 1997. Desde o lançamento do SixDegrees, houve centenas de outros SNSs, sempre com novas ferramentas, atraindo assim a audiência de novos usuários. Na Figura 2.1 é esquematizada uma linha do tempo relacionada à criação dos principais SNSs, sendo que o último mencionado é o MyChurch, lançado em 2006.

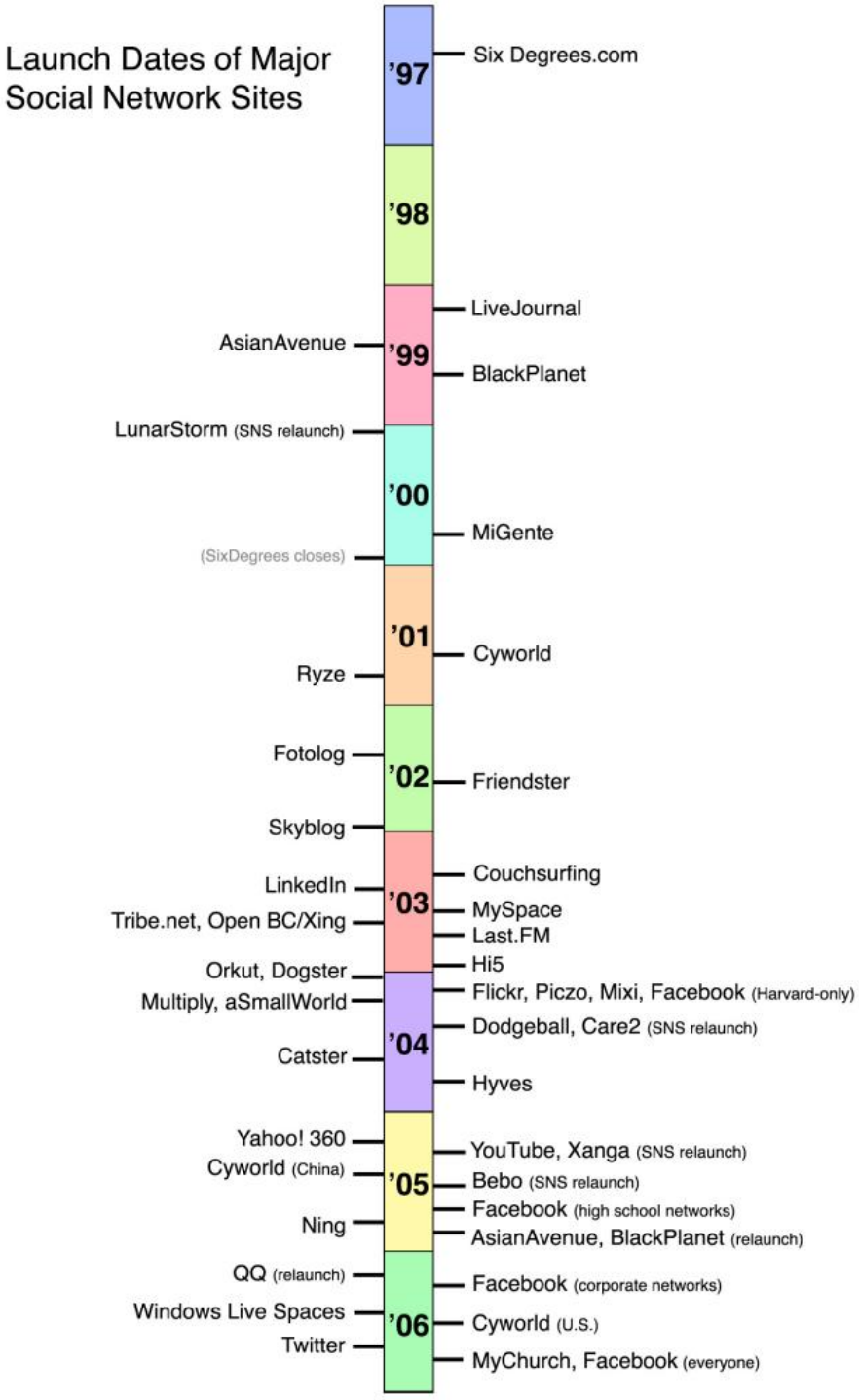


Figura 2. 1 Histórico de lançamento dos principais SNSs Boyd (2007).

Tanto o MiGente quanto o BlackPlanet, são SNSs voltados para grupos étnicos. O primeiro foi criado para latinos e o segundo para negros. Na Figura 2.2 é apresentada uma página do perfil de um usuário do MiGente. A foto do usuário é apresentada na área A e, na área B é exibida a lista de amigos do mesmo. Os outros componentes da página são informações complementares do perfil, como por exemplo, mensagem pessoal, artistas favoritos, informações pessoais, etc.

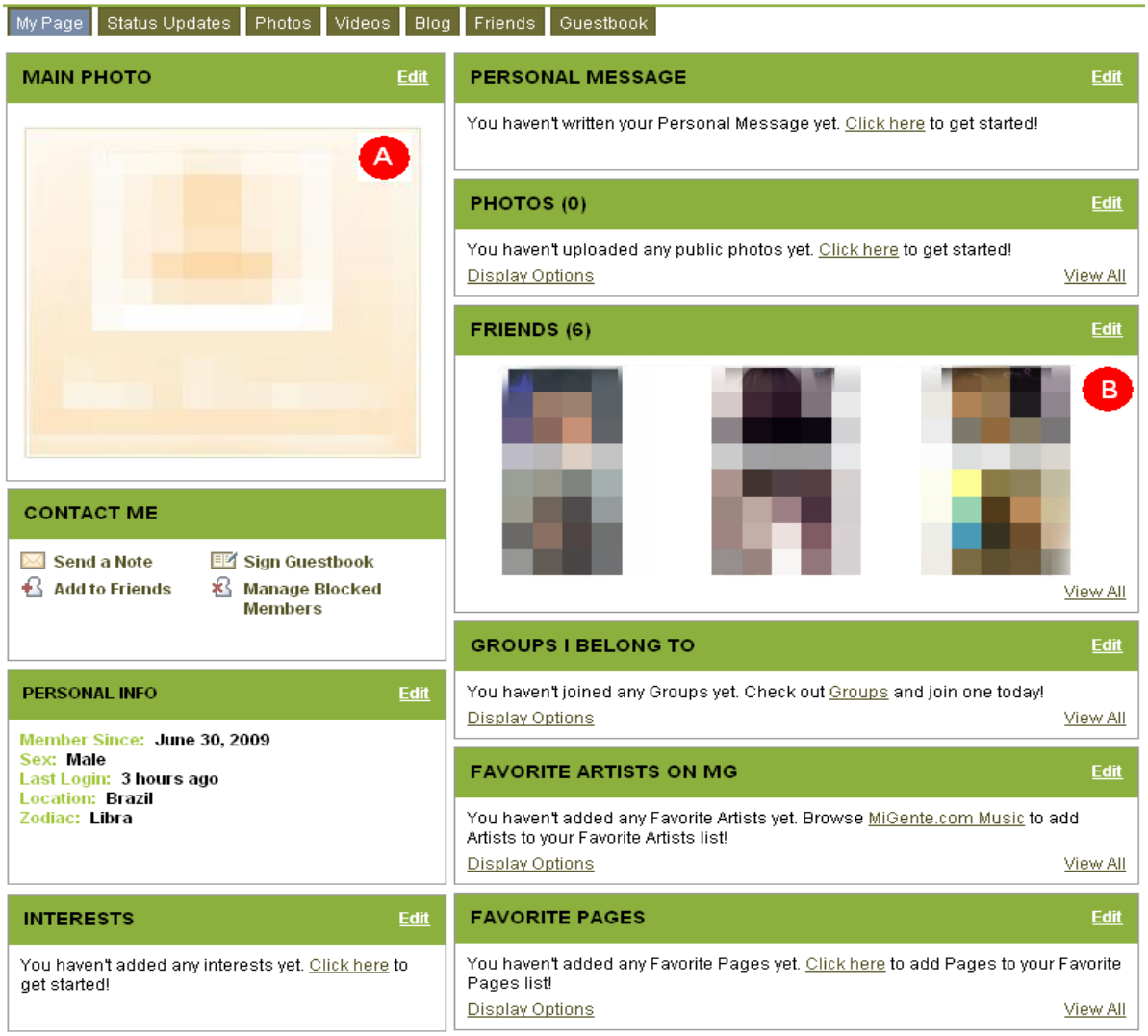


Figura 2. 2 Página do perfil de um usuário do SNS Migente.

SNSs como YouTube (www.youtube.com), Flickr (www.flickr.com), Delicious (www.delicious.com), Fotolog (www.fotolog.com), Last.FM (www.lastfm.com) e Photobucket (www.photobucket.com) são exemplos que tentam atrair a audiência de novos usuários com a possibilidade de construir redes sociais e compartilhar mídias, como vídeos, fotos, músicas e bookmarks1.

Outros SNSs, como o LinkedIn (www.linkedin.com), o Beehive (interno da IBM®) e o Peabirus (www.peabirus.com.br) são voltados para a vida profissional dos seus usuários. O primeiro permite que um usuário mantenha uma lista de contatos profissionais, no segundo, cada usuário é um funcionário de uma grande empresa e mantém sua lista de colegas

[1] Bookmarks são as páginas web favoritas de uma pessoa usuária de SNS.

de trabalho a fim de trocar experiências, e o terceiro é um SNS que permite que sejam construídas comunidades voltadas para negócios.

Um dos exemplos de SNS relacionados à educação é o LiveMocha (www.livemocha.com), que tem como objetivo unir pessoas que estão interessadas em aprender algum idioma estrangeiro , como por exemplo, inglês, francês, italiano, etc.

O Orkut, lançado em 2004, é o SNS mais popular entre os brasileiros. Atualmente mais de 50.60% dos usuários residem no Brasil, acompanhado de 20.44% na Índia e 17.78% nos Estados Unidos (fonte: www.orkut.com).

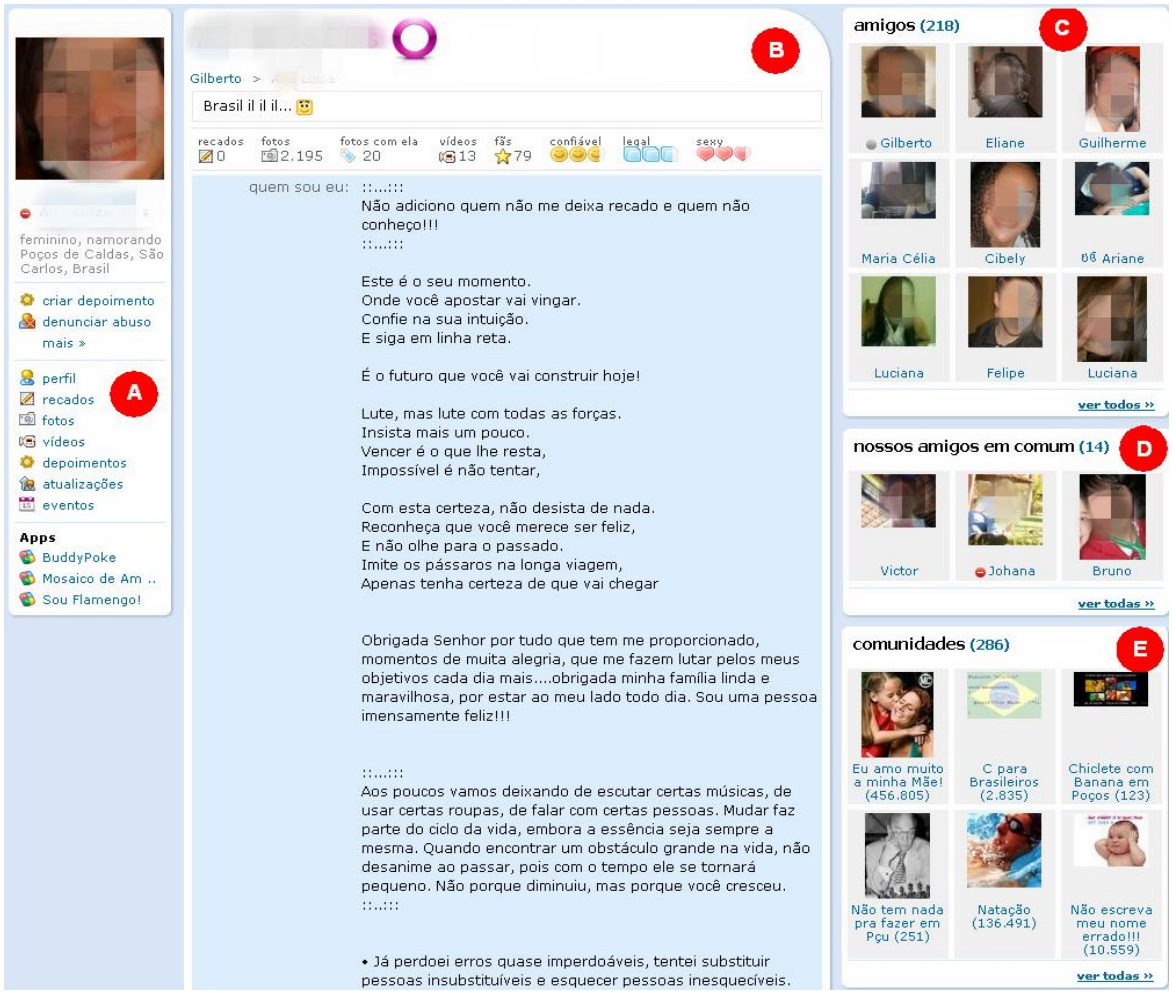


Figura 2. 3. Página do perfil de um usuário do SNS Orkut.

Na Figura 2.3 é apresentado o perfil de um usuário do Orkut. Na área A são exibidos, além dos dados pessoais do usuário, os vídeos favoritos, as fotos, os jogos, etc. Na área B é mostrada a segunda parte do perfil, isto é, quantas pessoas são fãs do usuário, o quanto as pessoas o consideram confiável e legal, e uma mensagem escrita pelo usuário, onde ele descreve sua personalidade (“quem sou eu”). Na área C é mostrada a lista de amigos,

como no MiGente e em qualquer outro SNS. Na área D são listados os amigos em comum entre o usuário que está visualizando o perfil e o usuário proprietário do perfil em questão. Finalmente, na área E, são mostradas as comunidades que o usuário participa.

O Orkut possui praticamente as mesmas características que os outros SNSs com o mesmo foco, ou seja, para o público em geral. Nesta lista, vale destacar o Hi5, sendo o mais usado no norte da América do Sul e América Central, o Sonico (www.sonico.com) que é o mais popular na América Latina e o Facebook (www.facebook.com), que depois de relançado em 2006, se tornou um fenômeno quanto à quantidade de usuários.

Finalmente, o Twitter (www.twitter.com) foi um dos últimos SNSs a serem lançados e conseguiu atrair milhões de usuários. Seu foco é voltado para a comunicação rápida entre os usuários. Ele utiliza o conceito de "seguidor" usado nos blogs, ou seja, ele permite que usuários sigam outros usuários a fim de receberem atualizações dos usuários "seguidos".

Na Tabela 2.1 são listados alguns dos SNSs da atualidade. Nessa tabela é mostrada a quantidade de usuários e o foco de cada um.

Tabela 2. 1 Relação dos principais SNSs ordenados pela quantidade de usuários.

SNS	Foco	Usuários[2]
Facebook	Geral. Compartilhamento de fotos, vídeos e bookmarks.	400.000.000
Qzone	Geral. Mais popular na China	200.000.000
Habbo	Geral. Sala de bate-papo e comunidades	162.000.000
Myspace	Geral. Mantém um site, com número limitado de página, para cada usuário.	130.000.000
Windows Live Spaces	Blogging, mensagens instantâneas.	120.000.000
Orkut	Geral. Compartilhamento de fotos e vídeos, e comunidades.	100.000.000
Friendster	Geral.	90.000.000
Hi5	Geral. Compartilhamento de fotos e vídeos, e comunidades.	80.000.000
Twitter	Micro-blogging	75.000.000

Ao observar a Tabela 2.1, é possível verificar o quanto os SNSs atraem as pessoas ao redor do mundo. Pela quantidade de usuários somados entre eles, não somente os mostrados na Tabela 2.1, mas também entre as centenas que existe, é possível afirmar que a maioria dos usuários está inscrito em mais de um SNS. Essa observação foi constatada no trabalho de Motoyama (2009).

[2] Fonte: http://en.wikipedia.org/wiki/List_of_social_networking_websites - julho de 2010.

2.3 A Busca por pessoas nos SNSs

O principal objetivo dos SNSs é facilitar ao usuário buscar e manter um maior número possível de contatos de amizades em torno de seus interesses, como por exemplo, hobbies, trabalho, visão política, religião, etc.

Unir as pessoas com interesses em comum tem como intuito tornar as redes sociais, gerenciadas pelos SNSs, mais ativas, pois permite ao usuário ter maior interesse no que os outros usuários estão escrevendo, nos assuntos abordados, etc., e por isso, acarretando em um maior uso e consequentemente maior produção de informação. É válido mencionar que essa informação poderia, por exemplo, ser uma poderosa ferramenta para estratégias de marketing.

Hoje em dia, para aumentar o número de contatos de amizade de um usuário, os SNSs disponibilizam serviços de buscas, recomendação de pessoas e, alguns, como é o caso do Sonico, Hi5 e Facebook, importações da lista de contatos de e-mail de um usuário.

A importação dos contatos de e-mail funciona da seguinte forma: o usuário informa quais são as suas contas de e-mail e as respectivas senhas. O aplicativo do SNS utiliza esses dados para acessar a lista de contatos do usuário nos respectivos serviços de e-mail, a fim de recuperar os e-mails dos contatos do mesmo. Em seguida, o aplicativo do SNS busca se algum dos contatos recuperados está cadastrado no SNS. Caso esteja, o contato é recomendado ao usuário para uma nova ligação de amizade.

Esse tipo de serviço dificulta aos usuários dos SNSs encontrar novos contatos de amizade, ele apenas faz com que um usuário transfira seus contatos de uma plataforma para outra, ou seja, não oferece ao usuário uma possibilidade de aumentar seus laços de amizade e/ou contatos para que ele possa ter um maior interesse em manter sua rede social mais ativa.

Os serviços de recomendação de pessoas, como por exemplo, o do Orkut e do Facebook trabalham sobre os contatos de amizade de um usuário. Eles utilizam o modelo de tríade discutido por Mislove *et. al.* (2008). Esse modelo verifica quem são os contatos de amizade de certo usuário, a partir desses contatos, o modelo idêntica quais são os seus outros contatos que o certo usuário não possui e os recomenda. Por exemplo, o usuário “Paulo” está conectado com a usuária “Maria”. Todos os contatos de “Maria” serão recomendados para “Paulo”, supondo que “Paulo” possa conhecê-los. Na Figura 2.4 é apresentado como esse tipo de serviço é usado no SNS Orkut.

Figura 2. 4.Sugestões de amizade do Orkut.

Esse tipo de serviço pode ter o mesmo resultado que o serviço de importação de e-mails, recomendando contatos de amizade já existentes na "vida real" de um usuário.

Os serviços de busca de usuários é outra forma que quase todos os SNSs têm para permitir ao usuário adicionar um outro em sua lista de contatos. Por exemplo, no Orkut, há a possibilidade de um usuário encontrar outro pelo nome, bem como permite que um usuário digite uma palavra-chave qualquer e encontre todos os usuários que usaram tal palavra em algum lugar em sua página de perfil ou Comunidade (discutida na seção 4.3.1).

Por exemplo, se um usuário está interessando em encontra um laço de amizade que goste da cidade do "Rio de Janeiro", ele poderia usar o nome da mesma como palavra-chave na busca. Isso faz com que ele tenha milhares de retornos, tanto oriundos das páginas de perfil dos usuários quanto das discussões nas Comunidades. A partir desses retornos, o usuário teria que fazer uma seleção entre milhares de usuários para verificar se alguns deles poderiam ter interesses comuns a ele em relação à cidade do "Rio de Janeiro".

Suponha também que esse usuário tenha interesse em encontrar outros usuários que estão falando sobre as belezas naturais do "Rio de Janeiro". Possivelmente ele pode encontrar, mas haverá um montante de usuários, selecionados pela busca, que estão falando sobre os problemas políticos, as olimpíadas, o tráfico de drogas e centenas de outros assuntos.

Um problema percebido nesse tipo de busca é que ela não consegue distinguir o interesse do usuário através do uso de palavras-chave. O usuário teria que usar várias palavras-chave na busca para tentar refinar um pouco mais os resultados, que pode não surtir efeito, pois a busca não considera a semântica envolvida entre as palavras, algo que poderia melhorar os resultados.

Além disso, esse tipo de busca não considera a cultura das pessoas, por exemplo, se houvesse essa flexibilidade, quando o usuário utilizasse a palavra-chave "Rio de

Janeiro" na busca, o serviço de busca poderia inferir que pessoas que utilizam as palavras "Cidade maravilhosa" e "Rio" podem também estar falando sobre o "Rio de Janeiro".

Esse problema acontece porque as buscas são realizadas apenas com comparações de palavras e não consideram um contexto maior, como por exemplo, as variações no vocabulário das pessoas, que pode variar de região para região, de país para país, ou seja, de cultura.

Existem alguns trabalhos, descritos na próxima seção, que têm como objetivo tentar aprimorar a forma com que os usuários podem identificar outros e consequentemente adicioná-los em sua lista de contatos. Alguns deles apresentam métodos que tentam identificar similaridades entre indivíduos, outros vão além, propondo a formação de redes sociais.

Esses trabalhos podem ser divididos em dois grupos, os que extraem redes sociais a partir de regras bem estabelecidas, e os que propõem a formação de redes sociais às pessoas. Independente de qual seja a categoria, todos, de uma forma ou de outra, partem de uma regra qualquer que infere similaridade entre os indivíduos.

2.4 Trabalhos relacionados com a identificação de pessoas que possam promover à aproximação social

Nos últimos anos, com o sucesso dos SNSs, surgiu um grande desafio de pesquisa, que está relacionado em definir o contexto entre as pessoas a partir de suas interações sociais ou uso da web, para identificar situações que podem promover um elo entre elas.

Um dos primeiros trabalhos nessa linha foi desenvolvido na década de 90 por Kautz *et. al.* (1997). Eles construíram um sistema (Referral Web) que é capaz de extrair redes sociais de especialistas. A partir de um nome de um especialista, cadastrado no sistema, são procurados outros nomes que coocorrem com ele em páginas web, em coautoria de artigos, em citações de artigos, em redes de notícias e organogramas (departamentos de universidades).

Nessa primeira década apareceram dezenas de outros novos trabalhos. Alguns deles conseguiram resultados expressivos que despertou o interesse pelos SNSs. Por exemplo, o Orkut, Hi5 e Facebook se baseiam na ontologia FOAF ("Friend of a Friend") de Finin (2005), ou seja, o que se baseou no modelo de tríade, discutido na seção anterior, para recomendar ligações entre as pessoas.

Essa ontologia armazena todos os amigos de um amigo de um usuário, para que eles sejam recomendados para o usuário em questão. As recomendações são feitas sobre a hipótese que a pessoa possa conhecer os amigos de seus amigos. Como exemplificado na seção anterior.

O que mais despertou interesse a partir da segunda metade dessa década foi a extração de redes sociais de especialista. O objetivo desses trabalhos eram facilitar o acesso aos especialistas dos mais variados assunto, de forma fácil e rápida. Alguns exemplos desses trabalhos são o de Matsuo *et. al.* (2006), Hamasaki *et. al.* (2006b), Hope (2006) e Tang *et. al.* (2008).

Matsuo *et. al.* (2006), utilizam duas técnicas diferentes para extrair redes sociais entre pesquisadores. Seus estudos de caso foram realizados em dois congressos: JSAI2003 (Japanese Society for Artificial Intelligence) e JSAI2004.

Na primeira técnica eles obtêm os nomes de todos os participantes dos congressos e os permutam formando pares. Para todos pares de nomes, utilizando maquinas de busca web, é procurado a coocorrencia de ambos em uma mesma página web. Caso isso aconteça é construída uma ligação entre eles.

A segunda técnica é usada para expandir a rede social conseguida anteriormente. Para isso cada participante do congresso recebeu um dispositivo que emite sinais. Um sensor capta esses sinais quando o individuo está próximo dele. Assim, quando um sensor recebe sinais de um grupo de pessoas dentro de um período de tempo, é formada uma rede social entre essas pessoas, considerando que elas estejam juntas.

O método que Hamasaki *et. al.* (2006b) e Hope (2006) propõem, também é aplicado em congressos científicos. Primeiramente uma mineração web é feita em busca de coocorrência dos nomes dos participantes do congresso em páginas web, em publicações, etc. Os nomes que coocorrem são ligados formando uma rede social.

Essa rede é expandida com a observação das interações sociais entre as pessoas durante o congresso, ou seja, as pessoas que conversam entre si são ligadas uma a outra na rede. Finalmente, a rede social extraída é apresentada aos participantes do congresso a fim deles adicionarem e/ou extraírem indivíduos.

O último exemplo de extração de redes sociais de especialistas é o de Tang *et. al.* (2008). Esse trabalho trata de extração de redes sociais acadêmicas relacionadas a tópicos de pesquisa. Por exemplo, podem ser formadas redes sociais em torno dos tópicos “IHC”,

“Engenharia de Software”, “Ergonomia”, etc. Para isso, são buscados os pesquisadores na DBLP (http://dblp.uni-trier.de/), que é um servidor de informações bibliográficas, e ligados um ao outro por meio do interesse de pesquisa.

Alguns trabalhos, como os de Jin *et. al.* (2007) e Joshi (2006) usam técnicas parecidas aos dos trabalhos anteriores, mas com o objetivos diferentes. O primeiro propõe a extração de redes sociais de atores contemporâneos. Eles assumem que dois atores podem ter uma ligação social, quando os nomes de ambos coocorrem em várias páginas web. O segundo se beneficia dessa mesma técnica, ou seja, ele vasculha a coocorrencia de nomes de celebridades e/ou autoridades nos artigos de notícias do Google News (http://news.google.com.br/), para criar uma ligação entre elas.

Bird *et. al.* (2006) e Tyler (2005) propõem extrair redes sociais de um serviço de e-mails. O método é relativamente simples, e é resumido da seguinte forma: quando uma pessoa recebe ou envia um email para outra pessoa, elas são ligadas devido a essa comunicação. A diferença é que Tyler (2005) considera um número maior de contato entre duas pessoas, possibilitando a extração de grupos mais acoplados para identificar situações que podem promover um elo entre elas.

Existem alguns trabalhos que se baseiam em outras especialidades para conseguir identificar situações que podem originar uma ligação entre duas pessoas. Por exemplo, o de Nunes (2008), que se baseia em teorias inerentes à psicologia para modelar, formalizar e armazenar um perfil psicológico de um usuário de um SNS, na qual é chamado de UPP (User Psychological Profile).

O UPP é gerado levando em consideração as respostas de um questionário. Ele pode ser usado pelo usuário a fim de identificar sua “Identidade Interna” ou por seus “amigos” com o objetivo de conseguir a “Identidade Social” dele.

Os resultados dos experimentos mostraram que o armazenamento e processamento do UPP do usuário por um sistema de recomendação pode prover uma recomendação com melhor qualidade, pois usuários com UPPs similares podem ser recomendados um ao outro. Porém, o grande desafio é encorajar usuários a responder o questionário extenso, composto por 900 questões, usado para gerar o UPP.

Finalmente, o trabalho de Chen *et. al.* (2009) usa as informações geradas pelas pessoas em SNSs, isto é, eles exploram uma regra relacionada ao vocabulário e ao interesse

delas, a saber: "*se nós postarmos conteúdo em tópicos similares, nós podemos estar interessados em conhecer um ao outro"* [3] (Chen, 2009, tradução nossa).

Para verificar esta hipótese eles desenvolveram um aplicativo que vasculha no SNS Beehive, quais as palavras que cada usuário mais utiliza em suas postagens e as mantêm em um "*saco de palavras*" [4] (Chen, 2009, tradução nossa). A recomendação de um usuário ao outro se dá a partir da similaridade entre seus "*saco de palavras*".

A pesquisa de Chen *et. al.* (2009) é a que mais se assemelha a proposta deste trabalho. Aqui também se usa as informações produzidas pelas pessoas e há a exploração do vocabulário das mesmas. A diferença é que neste trabalho é considerada a comparação semântica textual e a cultura das pessoas e no de Chen *et. al.* (2009) a semântica entre as palavras e a cultura das pessoas durante a comparação é ignorada.

2.5 Considerações finais

Nesse capítulo foi definido o que são os SNSs, mostrando, para isso, o histórico do lançamento e evolução dos principais sistemas desse tipo. Além disso, foram apresentadas algumas características e o foco que alguns deles possuem, como também alguns trabalhos que propõem técnicas para identificar pessoas que possam fazer parte de uma mesma rede de contatos.

Ao investigar os usuários do Orkut, Facebook, Hi5, Twitter entre outros, observou-se que eles, ou pelo menos boa parte, expõem seus conhecimentos, crenças, mitos, etc. nas suas interações sociais, ora por meio de troca de mensagens, ora por preenchimento de perfil, etc. Isso mostra que as pessoas consideram os SNSs como parte de seu dia a dia e se comportam neles como se estivessem na "vida real", sendo assim, para identificar usuários com interesses em comum é preciso mais do que comparar palavras inseridas por eles nos SNSs, pois a forma com que cada grupo de pessoas se expressa, mesmo falando de um mesmo assunto, pode ser diferente.

A partir do perfil do usuário e das suas interações sociais é possível saber do que ele gosta, qual seu ideal político, qual seu ponto de vista sobre um determinado assunto e muitas outras informações, ou seja, há uma exposição espontânea de sua cultura. Sob esta ótica pode-se dizer que os SNSs, além de serem considerados uma plataforma de informação do futuro Hope (2006), também podem ser considerados grande fonte de informação cultural,

[3] "if we both post content on similar topics, we might be interested in getting to know each other".
[4] "bag-of-words".

por isso, há a possibilidade de utilizar essa informação como apoio na identificação de pessoas que possam fazer parte de uma mesma rede de contatos.

É por esse motivo que este trabalho se propõe a definir um método que considera os valores, o conhecimento, as crenças, enfim, a cultura das pessoas com o intuito de identificar outras que possuam interesses em comum, mesmo se expressando de formas diferentes, para possivelmente fazerem parte de uma mesma rede.

Com o intuito de identificar a forma com que as pessoas se expressam, por exemplo, sobre um determinado assunto, considerando a cultura, este projeto faz uso de uma base que possui informações que representam o conhecimento cultural dos brasileiros. Essa base advinda do Projeto *Open Mind Common Sense* no Brasil (OMCS-Br) é descrita no próximo capítulo.

3 O PROJETO OPEN MIND COMMON SENSE NO BRASIL

3.1 Considerações iniciais

Neste Capítulo é descrito o projeto OMCS-Br (*Open Mind Common Sense no Brasil*), que tem como objetivo coletar e disponibilizar o conhecimento cultural dos brasileiros permitindo que esse conhecimento seja usado para contextualizar aplicações computacionais, ou seja, programas cujo conteúdo leva em consideração a cultura de um determinado grupo de pessoas, no caso, o grupo dos indivíduos brasileiros.

3.2 O Projeto OMCS-Br

Toda e qualquer pessoa adquire, ao longo de sua vida, conhecimento baseado nas suas experiências cotidianas. Esse conhecimento, mesmo que de forma não intencional, é compartilhado entre um determinado grupo social em uma determinada época e, por esse motivo, ele é chamado de conhecimento de senso comum ou conhecimento cultural, que é definido por Minsky como “as habilidades mentais que a maioria das pessoas compartilha” (MINSKY, 1986).

Nesse contexto, surgiram diversos projetos, como o *Open Mind Common Sense* (OMCS) (SINGH, 2002a) e (CYC) (LENAT, 1995), que visam coletar, processar e armazenar conhecimento cultural das pessoas a fim de poder utilizá-lo em aplicações computacionais. No Brasil, o LIA (*Laboratório de Interação Avançada*) da UFSCar, em colaboração com o *Media Lab* do MIT (*Massachusetts Institute of Technology*) (http://www.media.mit.edu/), lançou em agosto de 2005 o projeto OMCS-Br (TSUTSUMI, 2006).

Assim como no OMCS, no Projeto OMCS-Br considera-se que qualquer pessoa possui conhecimento cultural que possa prover às máquinas (SINGH, 2002b; SINGH, 2004), tornando a construção da base de conhecimento um trabalho colaborativo.

Na Figura 3.1 é apresentada a arquitetura do Projeto OMCS-Br. Na área verde é destacado o site utilizado para a coleta dos dados; na área azul, os elementos utilizados para processar o conhecimento coletado e; na área laranja, os elementos utilizados nas aplicações que utilizam conhecimento cultural.

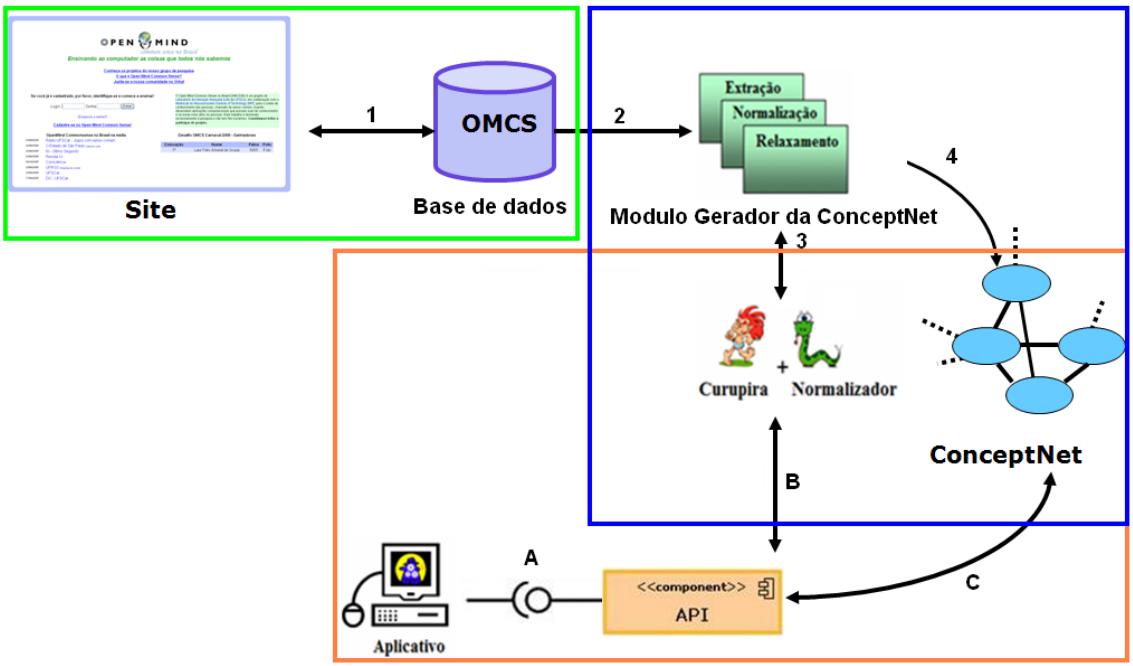


Figura 3. 1. Arquitetura do Projeto OMCS-Br.

Nas próximas seções são detalhadas cada uma dessas partes da arquitetura.

3.3 O Site

Para coletar o conhecimento cultural dos brasileiros, o Projeto OMCS-Br disponibiliza um site (destacado em verde na Figura 3.1) que pode ser acessado através do endereço http://www.sensocomum.ufscar.br.

Qualquer pessoa pode acessá-lo e se cadastrar para ter acesso a diversas atividades e temas disponíveis, tornando assim um colaborador do projeto. O site, até o momento, conta com oito temas distintos e vinte atividades, que abordam os diferentes tipos de conhecimento que compõem a cultura das pessoas (Figura 3.2).

Figura 3. 2. Temas e atividades que podem ser usadas pelo colaborador do Projeto OMCS-Br.

As atividades são usadas para coletar conhecimento cultural relacionados a assuntos gerais, como por exemplo, como os objetos são usados, onde são encontrados, quais suas propriedades, etc. Os temas são usados para coletar conhecimento cultural relacionado a certos domínios, como Sexualidade, Saúde, Universo Infantil, etc.

A coleta, tanto para os temas quanto para as atividades, é realizada por meio de *templates* (Figura 3.3), que são sentenças que possuem lacunas onde o colaborador preenche, de forma a compor uma sentença que, para ele, represente seu conhecimento cultural.

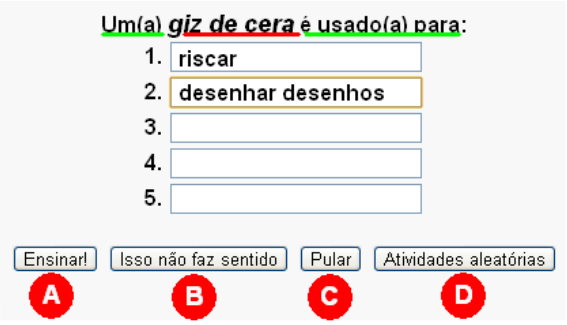


Figura 3. 3. Template para coleta de conhecimento cultural.

Os *templates* utilizados no site possuem uma parte estática e outra dinâmica. A parte estática (destacado em verde na Figura 3.3) é elaborada pelos mantenedores do Projeto. Já a dinâmica (destacada em vermelho na Figura 3.3) muda a cada vez que o *template* é apresentado ao colaborador, pois são consideradas as informações que outros colaboradores forneceram anteriormente, ou seja, há o uso de conhecimento já armazenado para obter novos conceitos relacionados.

Ao visualizar o *template* o colaborador pode executar quatro ações: a primeira, "Ensinar!" (A), deve ser utilizada para confirmar o preenchimento do *template,* isto é, enviar as informações para a base de dados do Projeto. A segunda, "Isso não faz sentido" (B), é uma alternativa para o colaborador informar ao gerador automático dos *templates* que o conteúdo da sentença, parte estática e dinâmica, não tem uma relação lógica. Isso fará que o gerador não use mais a formação do *template* em questão de novo. A terceira, "Pular" (C), faz com que o colaborador desconsidere o *template*. Por fim, a quarta, "Atividades aleatórias" (D), permite que *templates* de todas as atividades e temas existentes na base do projeto sejam utilizados, ou seja, o colaborador passa a receber *templates* de todas as atividades e temas de forma aleatória.

Após o preenchimento e o envio dos dados de um *template,* esses dados são armazenados na base de dados do Projeto. Essas sentenças passam por um processo de revisão, sendo que, após terem sido revisadas elas estarão preparadas para serem processadas de forma a gerar uma rede semântica. Esta rede é uma forma de representação de conhecimento, que pode ser vista como uma ferramenta de suporte para sistemas automatizados de inferência sobre o conhecimento (FALBO, 2004).

Na Figura 3.4 é apresentado um recorte dessa rede de conceitos, denominada ConceptNet (LIU, 2004). As informações coletadas através dos *templates* são os conceitos, representados pelos nós, interligados pelos arcos que representam a relação semântica entre os conceitos, que nesse projeto são as relações de Minsky.

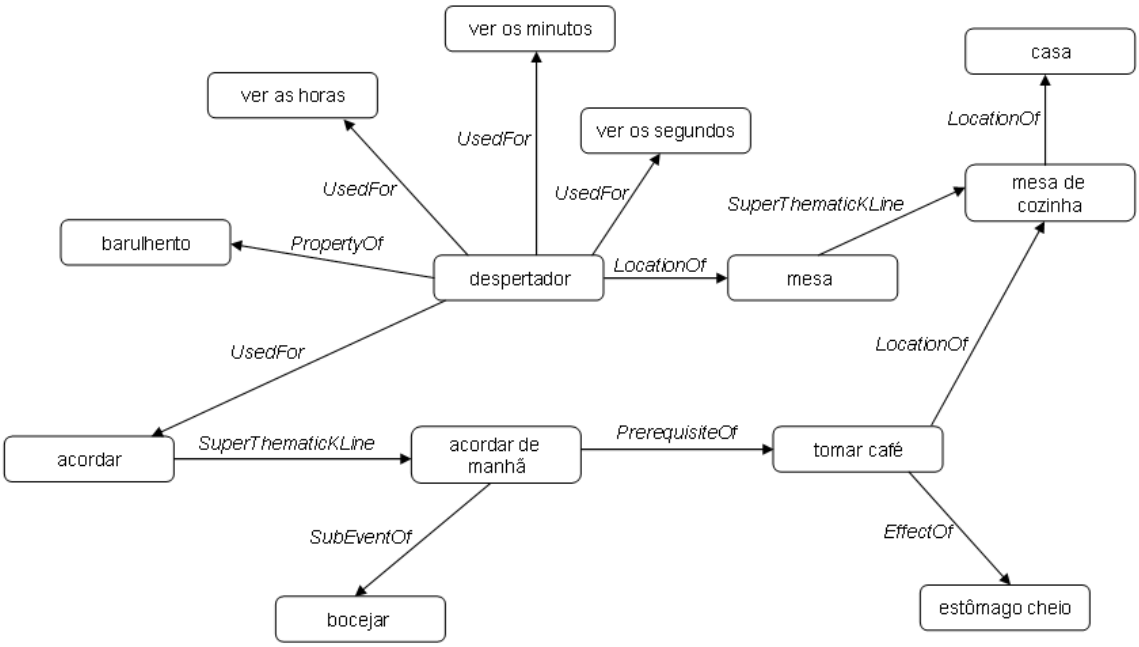


Figura 3. 4. Rede semântica – ConceptNet.

Atualmente, o Projeto utiliza 40 relações, sendo que 20 são negativas. Na Tabela 3.1 é apresentada uma tabela em que são mostradas todas as relações com exceção das negativas, pois as negativas são apenas prefixadas com o operador "Not", por exemplo: NotIsA.

Tabela 3. 1. Relações semânticas que compõem o projeto OMCS-Br (LIU, 2004).

Classe	Relação	Exemplo
K-Lines	ConceptuallyRelatedTo	(ConceptuallyRelatedTo 'bad breath' 'mint' 'f=4;i=0)
	ThematicKLine	(ThematicKLine 'weding dress' 'veil' 'f=9;i=0)
	SuperThematicKLine	(SuperThematicKLine 'western civilisation' 'civilisation' 'f=0;i=12)
Things	IsA	(IsA 'horse' 'animal' 'f=17;i=3)
	PropertyOf	(PropertyOf 'fire' 'dangerous' 'f=17;i=1)
	PartOf	(PartOf 'butterfly' 'wing' 'f=3;i=0)
	MadeOf	(MadeOf 'bacon' 'pig' 'f=3;i=0)
	DefinedAs	(DefinedAs 'meat' 'flesh of animal' 'f=2;i=1)
Agents	CapableOf	(CapableOf 'dentist' 'pull tooth' 'f=4;i=0)
Events	PrerequisiteEventOf	(PrerequisiteEventOf 'read letter' 'open envelope' 'f=2;i=0)
	FirstSubeventOf	(FirstSubEventOf 'start fire' 'light match' 'f=2;i=3)
	SubEventOf	(SubEventOf 'play sport' 'score goal' 'f=2;i=0)
	LastSubeventOf	(LastSubEventOf 'attend classical concert' 'aplaud' 'f=2;i=1)
Spatial	LocationOf	(LocationOf 'army' 'in war' 'f=3;i=0)
Causal	EffectOf	(EffectOf 'view video' 'entertainment' 'f=2;i=0)
	DesirousEffectOf	(DesirousEffectOf 'sweat' 'take a shower' 'f=3;i=1)
Functional	UsedFor	(UsedFor 'fire place' 'burn' 'f=1;i=2)
	CapableOfReceivingAction	(CapableOfReceivingAction 'drink' 'serve' 'f=0;i=14)

Affective	MotivationOf	(MotivationOf ‘play game’ ‘compete’ ‘f=3;i=0)
	DesireOf	(DesireOf ‘person’ ‘not be depressed’ ‘f=2;i=0)

As relações são separadas por classes, sendo que cada uma dessas classes representa um determinado domínio. Por exemplo, a classe *Events* (veja Tabela 3.1) possui quatro relações semânticas, das quais conseguem representar tudo, ou quase tudo, sobre eventos. *PrerequisiteEventOf* representa um evento obrigatório antes de outro evento. *FirstSubeventOf* especifica o primeiro evento, entre um conjunto de eventos, que leva a certo evento. *SubEventeOf* especifica o evento que acontece depois de certo evento. Finalmente, *LastSubeventOf* aponta o último evento, entre um conjunto de eventos, relacionado a um primeiro evento. Um estudo completo de todas as relações que foi feito por Gilberto Astolfi, Johana Maria Rosas Villena e David Buzatto. Esse estudo pode ser visto no Apêndice A.

3.4 A geração da ConceptNet

O modulo gerador da ConcepNet (destacado em azul na Figura 3.1), é composto por três fases: Extração, Normalização e Relaxamento.

Na Fase de **Extração**, cada sentença armazenada na base de dados é submetida a regras que as dividem em fragmentos que irão compor um nó (conceito) da rede semântica. Além disso, para ligar semanticamente tais fragmentos, é atribuído uma das relações de Minsky. Esta relação não é escolhida aleatoriamente, quando um *template* é construído uma ou mais relações são associadas a ele. Por exemplo, o *template* da Figura 3.3 está associado explicitamente, mas não para o colaborador, à relação *UsedFor*. Nesse caso, com a formação da sentença “Um(a) **giz de cera** é usado(a) para **desenhar desenhos**”, após a fase de extração será gerado “UsedFor ‘giz de cera’ ‘desenhar desenho’”, como apresentado na Figura 3.5.

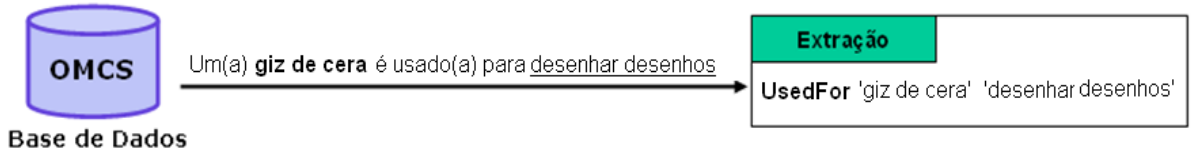


Figura 3. 5. Fase de extração.

***A Fase de* Normalização** é apoiada pelo Curupira (MARTINS, 2002), um analisador sintático para o português brasileiro. Ele se encarrega de categorizar cada uma das palavras, ou seja, etiquetá-las como verbo, ou substantivo, ou artigo, etc. A normalização se concretiza pelo fato de colocar, com o apoio do dicionário Delaf (MUNIZ, 2004), os conceitos em sua forma canônica, ou seja, os substantivos e os adjetivos são colocados no singular, os verbos no infinitivo, etc. (Figura 3.6).

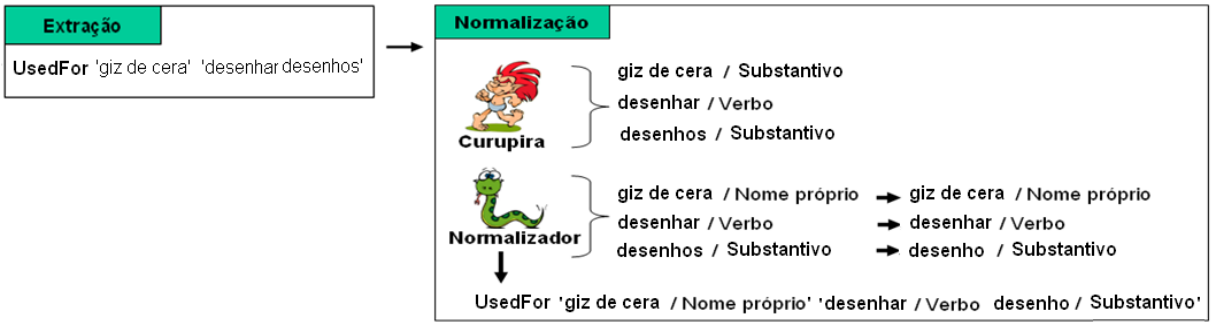


Figura 3. 6. Fase de Normalização.

A Fase de **Relaxamento** consiste em inserir os metadados "*f*" e "*i*" na relação. O "*f*" significa a frequência que uma mesma relação foi construída a partir de um conhecimento provido pelos usuários. Por exemplo, se dois usuários preencherem o *template* da Figura 3.3 com as mesmas informações, isto é, "Um(a) **giz de cera** é usado(a) para **desenhar desenhos**", este fato de conhecimento cultural após processado terá f = 2 (Figura 3.7).

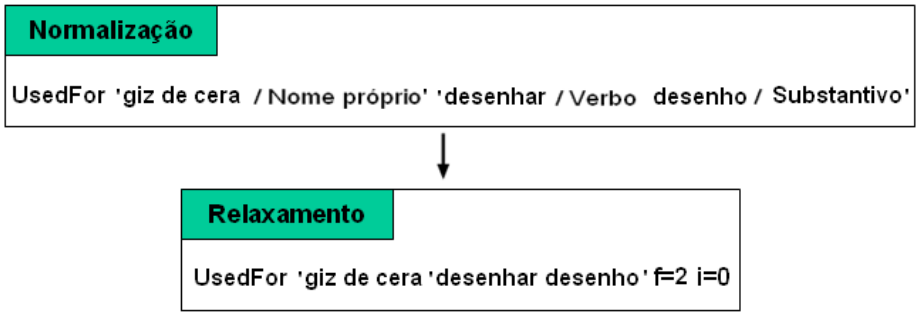


Figura 3. 7. Fase de Relaxamento.

O "*i*" – de inferência – significa a quantidade de vezes que uma relação foi gerada a partir de regras de inferência, que permitem que novas relações sejam geradas automaticamente a partir de relações já existentes (SILVA, 2009b). Por exemplo, um usuário preenche um *template* que gera a seguinte sentença: "O mar é composto de ondas fortes". Percebe-se que "ondas fortes" é um trecho de texto com um substantivo seguido de um adjetivo. Para esse caso há uma regra que gera a relação "PropertyOf 'onda' 'forte' f=0 i=1". Além da relação que já estava associada ao *template*, que no exemplo é "MadeOf 'mar' 'onda forte' f=1 i=0".

Com o fim de todo o processamento sequencial de extração, normalização e relaxamento é então gerada a ConceptNet (Figura 3.4).

Na próxima seção é apresentado como o perfil dos colaboradores do Pojeto OMCS-Br são associados a um fato de conhecimento.

3.5 O perfil dos colaboradores

Quando um colaborador se cadastra no site do OMCS-Br, ele deve fornecer algumas informações como faixa etária, sexo, área de interesse, país de origem, o estado e a cidade onde reside. Essas informações são utilizadas, posteriormente, para filtrar contribuições, ou seja, conhecimento cultural de acordo com os perfis de colaboradores desejados.

Por esse motivo, os fatos de conhecimento cultural provido por um colaborador é armazenado juntamente com os dados de seu perfil. Nas seções anteriores não foi tratado os exemplos das relações semânticas com todos os seus dados para mostrar o processo sendo construído passo a passo, mas deixa-se claro que o resultado final do processo, desde a sentença fornecida pelo colaborador até a formação de uma relação semântica, o conhecimento fica armazenado como é apresentado na Figura 3.8.

```
(UsedFor "computador" "estudar" "M" "18_29" "mestrado" "Clementina" "SP" "1")
(UsedFor "computador" "jogar" "M" "18_29" "mestrado" "Clementina" "SP" "2")
(UsedFor "computador" "trabalhar" "M" "18_29" "mestrado" "Clementina" "SP" "3")
(UsedFor "arquivo" "agrupar dados" "F" "18_29" "mestrado" "São Carlos" "SP" "8")
(UsedFor "cadeira" "sentar" "M" "18_29" "mestrado" "Clementina" "SP" "9")
(UsedFor "arquivo" "guardar informações" "M" "18_29" "mestrado" "Clementina" "SP" "10")
(UsedFor "computador" "pesquisar" "M" "18_29" "mestrado" "Clementina" "SP" "11")
(LocationOf "computador" "laboratório de informática" "M" "18_29" "mestrado"
"Clementina" "SP" "12")
(LocationOf "sofá" "sala de estar" "M" "18_29" "mestrado" "Clementina" "SP" "13")
```

Figura 3. 8. Exemplo de associação entre a relação semântica e o perfil do colaborador.

Portanto, cada fato de conhecimento cultural é composto pelos seguintes atributos: o primeiro argumento de cada fato refere-se à relação de Minsky; o segundo e o terceiro, aos conceitos; do quarto ao oitavo são organizadas as informações de perfil e; o nono, ao *id* da sentença armazenada no banco de dados, permitindo assim que por meio de um determinado fato, consiga-se obter a sentença original, que é armazenada em linguagem natural.

3.6 Acesso a ConceptNet

O acesso à ConceptNet (destacado em verde na Figura 3.1) é provido por uma API (*Application Programming Interface*). Uma das funções disponibilizadas é a *getContext*. Ela recebe um conceito como parâmetro que pode ser um substantivo, um verbo ou um adjetivo já na forma canônica (CARVALHO, 2007), e retorna todo o conhecimento relacionado ao conceito pesquisado. Na Figura 3.9 é apresentado um exemplo de aplicativo

que utiliza esta função. Nesse caso o conceito utilizado para a busca foi “Giz” (área A da Figura 3.9). Na área B da Figura 3.9 é apresentado o retorno da API e na área C a representação gráfica do retorno.

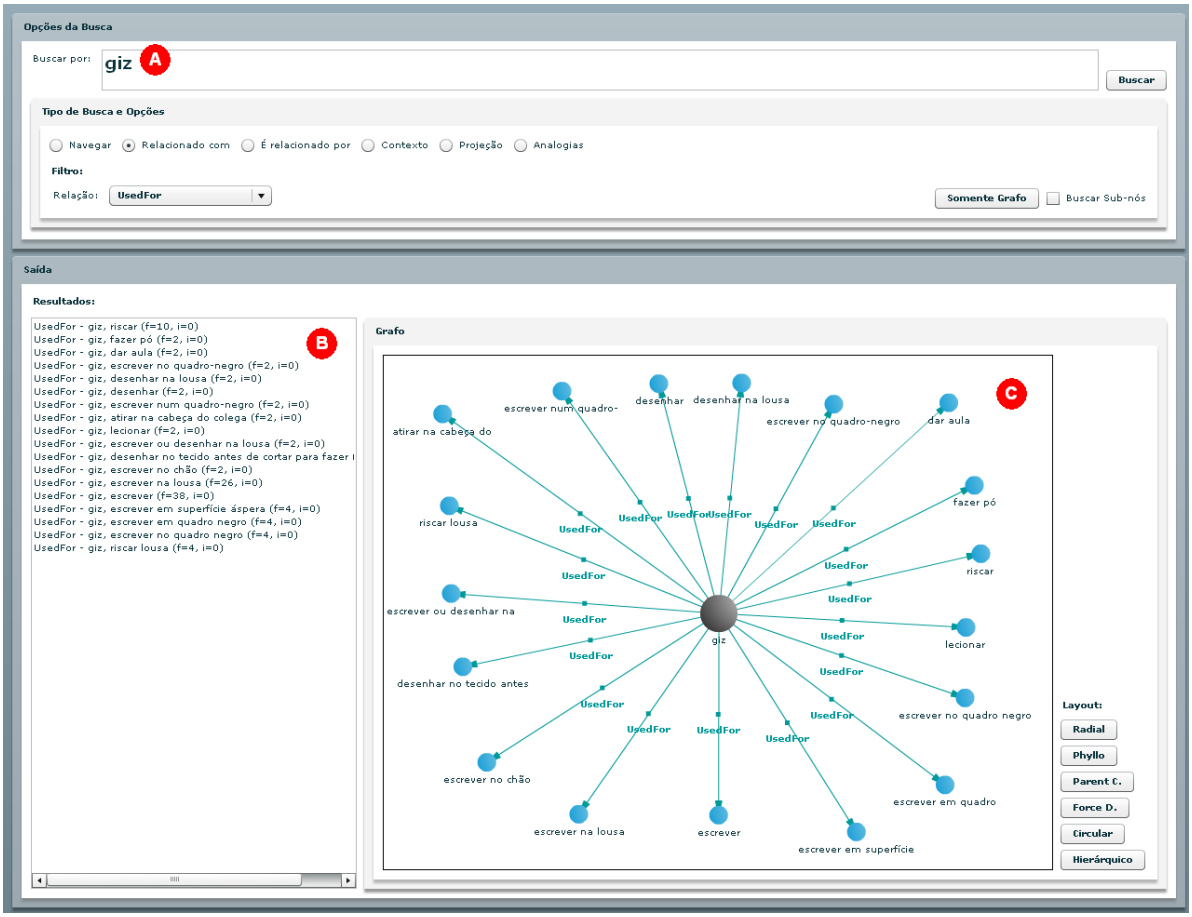


Figura 3. 9. Exemplo de uso da API da ConceptNet. Esse aplicativo foi desenvolvido por David Buzatto.

Neste trabalho é considerado apenas a função *getContext()*, pois é a única que é utilizada, mas existem outras que estão relacionadas a analogias, contextos, projeções, etc.

3.7 Considerações finais

Nesse capítulo foi apresentado o Projeto OMCS-Br, detalhando como o conhecimento cultural da população brasileira é coletado, processado e disponibilizado às aplicações computacionais que utilizam esse conhecimento. Essas explicações são de grande importância para o entendimento geral do trabalho, pois o mesmo usa de forma indispensável o conhecimento cultural mantido pela base do OMCS-Br.

4 ETAPAS PARA A OBTENÇÃO DO MÉTODO

4.1 Considerações iniciais

Este capítulo descreve todas as iterações realizadas para a obtenção do método de identificação de pessoas, usuárias de SNSs, que estão falando sobre um mesmo assunto considerando as informações culturais. Cada uma se baseia fielmente na abordagem de trabalho proposta na seção 1.4.

Na iteração um, discutida na seção 4.2, é descrita a primeira versão do método. Essa versão objetiva identificar pessoas que estão falando sobre o mesmo assunto, considerando o perfil e o conhecimento cultural das mesmas; pois nessa fase acredita-se que pessoas que possuem o mesmo perfil possivelmente podem possuir um vocabulário comum, quando se trata de um mesmo assunto.

Na seção 4.3 é apresentada a segunda versão do método, que foi instanciada em uma nova iteração da abordagem de trabalho proposta. Essa versão considera somente o vocabulário das pessoas, independente de seu perfil, como uma forma de reuni-las em torno de um assunto.

A terceira versão do método é instanciada na iteração três e discutida na seção 4.3. Ela, além de considerar um vocabulário comum entre em pessoas, como um requisito para reuni-las, também considera o contexto[5] no qual o assunto que está sendo discutido se insere. Essa evolução foi essencial para que se pudesse sugerir que as pessoas identificadas, além de estarem falando sobre o mesmo assunto, podem se interessar pelo mesmo.

Finalmente, na seção 4.4 é apresentada a última versão do método, instanciada na iteração quatro. Ela é o produto deste trabalho e sua evolução a capacitou em identificar pessoas, que estão falando sobre o mesmo assunto e contexto, considerando diferenças culturais de cada indivíduo, ou seja, considerando as diferentes formas como as pessoas se expressão sobre um determinado assunto.

Todas as versões do método se beneficia de alguma forma da base de conhecimento cultural do projeto OMCS-Br. A medida que o método foi se tornando cada vez mais maduro, mais a dependência de conhecimento cultural foi aumentando, tornando-se parte indispensável para o método.

---

[5] O termo contexto referido por este trabalho está relacionado com a situação e circunstancia em que um texto ocorre, como o tempo, cultura e a forma de expressão. Um exemplo é o uso da palavra “alavancar”, que quando usada no contexto da economia refere-se ao aumento de preços, ações e outras coisas relacionadas, em outro contexto, como o da construção civil, ela pode ser usada quando algo pode ser levantado ou suspendido do chão.

A seguir são descritas em detalhes todas as versões do método instanciadas sobre a metodologia de trabalho proposta.

4.2 Iteração 1 – Identificando pessoas com o mesmo perfil e vocabulário

Essa iteração se concentra em usar o perfil e o vocabulário dos colaboradores do projeto OMCS-Br, a fim de identificar pessoas em SNSs que estão falando sobre o mesmo assunto. Para isso acredita-se que pessoas que possuem o mesmo perfil, que nesta iteração está relacionado a pessoas com a mesma idade, sexo, local onde mora e grau de escolaridade, certamente possuem o mesmo vocabulário quando falam sobre um mesmo assunto.

Inicialmente há a necessidade de definir um assunto, representado por uma única palavra. Tal assunto é submetido à base de conhecimento cultural, a fim de recuperar um conjunto de conceitos relacionados culturalmente a esse assunto e, um perfil que represente esse vocabulário.

Esses dados são utilizados na busca por pessoas em SNSs que possuem o mesmo perfil e que fazem uso do vocabulário, obtido da base, quando falam sobre o assunto em questão. Com isso são agrupadas pessoas em torno de um mesmo assunto e que possivelmente poderiam ter chances de se interessarem por ele.

4.2.1 Exposição do problema

Existem alguns trabalhos na literatura, discutidos no capítulo 2, que propõem identificar pessoas dentro de um mesmo contexto que possa promover um elo entre elas (FIGUEIRA FILHO, 2008). Tais contextos são procurados ora em SNSs, ora na web, ora em congressos, etc.

Esses contextos são definidos por regras, como por exemplo, a coocorrência de nomes de pessoas nas mesmas páginas web, ou nos mesmos congressos, ou na possibilidade de pessoas compartilharem o mesmo vocabulário em SNSs, etc.

O trabalho de Chen (2009), por exemplo, explora a regra relacionada ao vocabulário e ao interesse das pessoas. A confirmação que duas pessoas têm os mesmos interesses é obtida pela medida de similaridade entre seus "*saco de palavras*" (*cf. Cap. 2*). Outro exemplo é o trabalho de Nunes (2008), a regra definida por eles para inferir que duas pessoas são similares é baseada na igualdade dos *UPPs* de ambos (*cf. Cap. 2*).

Esses dois trabalhos, como os demais apresentados no capítulo 2, assumem regras que podem falhar. Por exemplo, não se pode afirmar que pessoas possuem os mesmos

interesses apenas por utilizarem as mesmas palavras, pois elas podem estar sendo usadas em assuntos e/ou contextos completamente diferentes. Além disso, a medida do interesse de uma pessoa é algo tão subjetivo, que não se pode inferir seu interesse por um assunto apenas por ela usar algumas palavras inerentes a ele. Também o fato de duas pessoas possuírem o mesmo "perfil psicológico" (*cf. Cap. 2*) não implica que elas tendem a possuir os mesmos interesses, independente do assunto em questão.

Esse é o problema que o método proposto por este trabalho pretende amenizar, pois o mesmo tem como intuito identificar grupos de pessoas que estão falando sobre o mesmo assunto em SNS, para poder inferir, com menos chances de erro, que tais grupos podem se interessar pelo assunto no qual foi usado como contexto de busca.

4.2.2 Resolução do problema

O método proposto aqui se inspira nas abordagens de Chen (2009) e Nunes (2008), pois procura explorar o vocabulário e o perfil das pessoas, a fim de identificar situações onde ambos possam auxiliar na definição de uma regra que seja capaz de identificar pessoas que estão falando sobre o mesmo assunto.

Com base nos trabalhos citados anteriormente, em um primeiro momento assumiu-se a seguinte regra:

(1) ***pessoas que possuem o mesmo perfil podem utilizar o mesmo vocabulário quando se expressam sobre um mesmo assunto.***

É válido mencionar que essa regra evoluiu à medida que novos caminhos em busca do objetivo do trabalho foram sendo percorridos.

O perfil a que essa regra se refere não é o mesmo proposto por Nunes (2008), mas um perfil que especifica a idade, o sexo, a localização e a escolaridade de uma pessoa. Já o vocabulário é referente a conceitos relacionados a um assunto qualquer.

Por exemplo, uma instância para um perfil referente ao assunto "política" poderia ter os seguintes dados: idade = "*acima de 60 anos*"; sexo = "*masculino*"; cidade = "*São Carlos*"; estado = "*São Paulo*"; escolaridade = "*superior completo*". O vocabulário poderia ter as seguintes palavras: "*Senado, senador, presidente, Lula, Dilma, Serra, deputado* e *eleições*".

O ponto que se pretende chegar com essa abordagem é uma associação entre os dados do perfil e do vocabulário, assim, poderia ter indício de que um determinado perfil,

quando se refere ao assunto "política" por exemplo, usa tal conjunto de palavras como seu vocabulário específico.

Nessa abordagem as palavras do vocabulário caracterizam sempre um assunto. Diferentemente de Chen (2009), que pode identificar um "*saco de palavras*" para um indivíduo com um vocabulário tão heterogêneo que seria impossível identificar o interesse do mesmo. O exemplo do "*saco de palavras*" de um indivíduo apresentado por ele deixa claro esta afirmação: "*janeiro, ofício, pessoas, boston, encontrar, Roma, pai, halloween e mestre*"[6] (Chen, 2009, tradução nossa).

Além disso, de acordo com o exemplo anterior, não é possível afirmar que as pessoas com os "*sacos de palavras*" idênticos estão falando sobre o mesmo assunto, pois devido à heterogeneidade das palavras, elas podem ter sido oriundas de assuntos completamente distintos, em contextos completamente diferentes. Isso mostra a subjetividade em definir o interesse de uma pessoa.

Por outro lado, quando se considera apenas um assunto, pelo menos parte desse problema pode ser resolvido, porque percebe-se que para cada assunto há, possivelmente, um número específico de palavras que sempre coocorrem, fazendo com que a representação do vocabulário seja mais homogêneo em relação ao tema.

É o que se pretende quando aqui é proposto o uso de um perfil associado a um conjunto de palavras referente a um assunto, que a partir deste ponto será chamado de *PC* (*Perfil Cultural*). *PC* pode ser visto como uma variável que é usada como parâmetro em uma busca em certo SNS, a fim de identificar pessoas que possuem perfil e vocabulário similares a ela. Um exemplo de *PC* pode ser o seguinte: pessoas de sexo masculino, com idade entre 18 e 29 anos, residente no estado de São Paulo, na cidade de Clementina, com o nível de Mestrado e que usa as palavras "senado", "população", "governo" e "povo" quando se referem ao assunto política.

A construção de *PC* é apoiada pela base de conhecimento cultural do projeto OMCS-Br. A justificativa para isso se dá pelo fato da base apresentar indícios de ser capaz de representar, por amostragem (SILVA, 2009b), o que um grupo de pessoas, com um determinado perfil, têm como conhecimento cultural em relação a certo assunto.

Alguns trabalhos, como por exemplo, Ferreira (2008) e Silva (2009a) utilizam a base do OMCS-Br para recuperar conhecimento cultural considerando um perfil. O processo

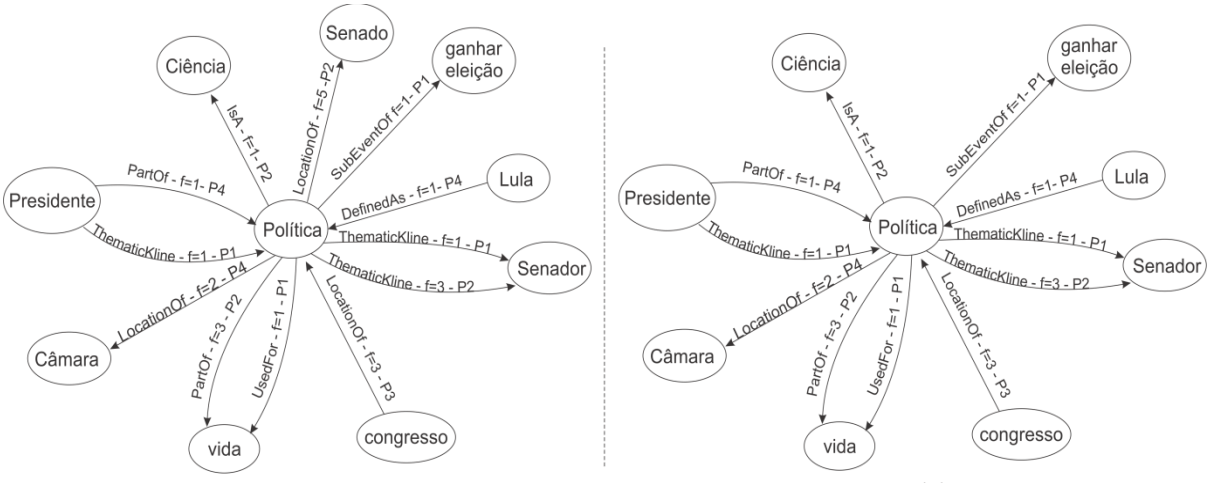
[6] "January, craft, people, boston, meet, rome, dad, halloween and master".

é realizado da seguinte maneira: são submetidos alguns dados, tais como idade, escolaridade, sexo, que representem um perfil à base, a fim de recuperar somente o conhecimento provido por colaboradores com o mesmo perfil. Com isso, eles conseguem contextualizar suas aplicações de acordo com o perfil dos usuários.

Para a construção de *PC* é realizado o inverso, ao invés de submeter um perfil em busca de conhecimento, é submetido um assunto em busca de um perfil e vocabulário. A seguir é apresentada a descrição de como o método funciona para construir o *PC*.

Inicialmente o método recebe como entrada um assunto, que pode ser obtido no perfil de um usuário de SNS, representado por um conceito (uma palavra ou expressão). Esse assunto é usado para extrair os conceitos relacionados na base do OMCS-Br, ou seja, o conhecimento relacionado ao assunto.

Todo o conhecimento obtido é representado por uma rede semântica, que além de manter o conhecimento relacionado ao assunto, também mantém, implicitamente por meio de referência, o perfil de cada colaborador associado ao conceito, uma vez que é possível identificar quem inseriu esse conceito. Na Figura 4.1 é ilustrado um exemplo usando o conceito "política".



Legenda:
P1 = "M", "13_17", "Ensino médio incompleto", ''São Carlos", "SP"
P2 = "M", "18_29", "Mestrado", ''Clementina", "SP"
P3 = "F", "18_29", "Mestrado", ''São Carlos", "SP"
P4 = "M", "30_45", "Ensino médio completo", ''São Carlos", "SP"

Figura 4. 1. (a) parte dos conceitos obtidos por meio da busca na base do OMCS-Br com o perfil dos colaboradores. (b) estado da rede semântica depois de extraído os conceitos com maior frequência.

É feita uma busca na rede semântica a fim de identificar os conceitos, relacionados ao assunto, que mais se repetem, isto é, os conceitos que possivelmente estão mais ligados semanticamente com o assunto, que nesse exemplo é "política".

Por exemplo, no caso da Figura 4.1 (a) é o conceito “senado”, pois sua frequência é 5 (f=5), a maior entre todos os outros. Esse conceito, assim como os outros de maior frequência, é extraído da rede semântica juntamente com perfil correspondente, e armazenado em um vetor[7], diminuindo a rede existente na Figura 4.1 (a). Na Figura 4.1 (b) é mostrado como a rede semântica fica depois do procedimento, observe que o conceito “senado” foi extraído.

O processo é repetido novamente até a rede semântica ficar com a metade dos conceitos do seu tamanho original (essa condição de parada é apenas para experimentação inicial do método, podendo ser melhor controlada caso essa solução tenha bons resultados). Por exemplo, nesse caso a rede não pode ficar com menos de cinco conceitos, com exceção do conceito usado na busca, que nesse caso é “política”. Na Figura 4.2 é mostrado como a rede fica no final do processo, ou seja, com os conceitos com maior frequência extraídos.

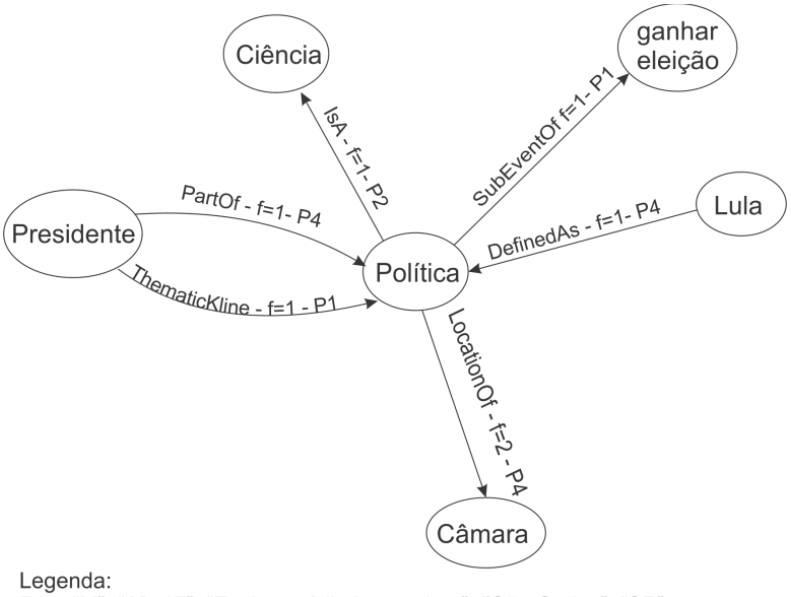


Figura 4. 2. Estado da rede semântica depois dos conceitos com maior frequência serem retirados.

O vetor com os conceitos e os perfis obtidos da rede semântica é analisado com o intuito de verificar o perfil que mais se repete, a fim de elegê-lo como o perfil que representa o vocabulário formado pelos conceitos recuperados da rede semântica no processo anterior. Na Figura 4.3 é ilustrado esse processo.

[7] Estrutura de dados usada na computação (TENENBAUM, 1995).

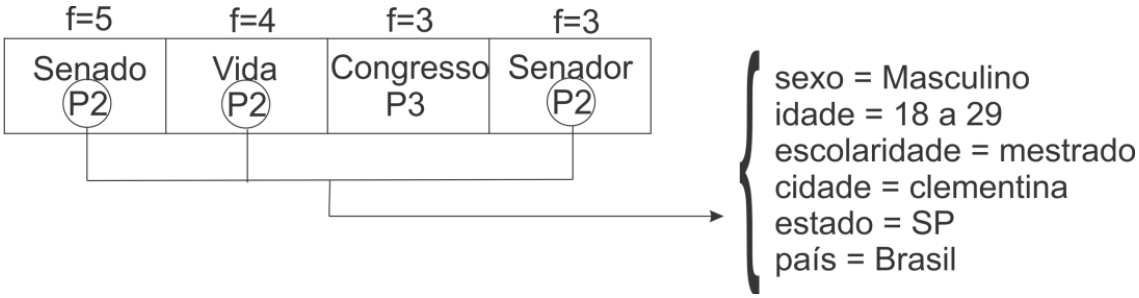


Figura 4. 3. Simulação da busca do perfil que mais representa certo vocabulário. Conceitos com maior frequência retirados da rede semântica.

Finalmente tem-se um vocabulário, formado por um conjunto de palavras e um perfil, isto é, *PC*, que é formalizado da seguinte forma: *PC (idade, sexo, escolaridade, país, estado, cidade, concepts [c_1, ..., c_n]) – onde c_n* é uma palavra pertencente ao vocabulário.

Com isso assume-se que as pessoas, usuárias de SNS, que têm perfil semelhante a *PC* provavelmente tendem a se expressar com algumas das palavras armazenadas pelo vetor *concepts* quando se referem ao assunto em questão. É válido mencionar que a ideia de *PC* surgiu com base em dois trabalhos, uma parte da estrutura de dados que o representa o perfil, *idade, sexo, escolaridade, país, estado e cidade*, é inspirada no trabalho Nunes (2008). A outra parte, *concepts [c_1, ..., c_n]*, é inspirada no trabalho de Chen (2009) que considera o vocabulário das pessoas. As razões para isso é que Chen (2009) consegue representar o que as pessoas falam nos SNSs, e Nunes (2008) consegue representar quem são estas pessoas.

Tabela 4. 1. Exemplo de perfil extraído de certo SNS.

Nome	Idade	Sexo	País	Estado	Cidade	Escolaridade	words
João Souza	19	M	Brasil	SP	Clementina	Mestrado	vida, Brasília, amor, cidade, flor, Senador, natural, Sarney, população, time, congresso, eleição, Senadora, etc.

A segunda parte do método visa procurar em um SNS perfis, como o exemplo da Tabela 4.1, semelhantes a *PC*. Esses perfis são representados por *P (nome, idade, sexo, escolaridade, país, estado, cidade, words [w_1, ..., w_n])*, onde w_n é uma palavra qualquer que o usuário do SNS usou em seu perfil, e os demais dados são referentes a localização, idade, etc. que compõe o perfil de um usuário em um SNS, como é mostrado no exemplo do SNS Orkut na Figura 2.3 área A apresentada no capítulo 2.

A implicação que *P ~ PC* é satisfeita pelas seguintes verificações:

(1) Se os campos idade, sexo, escolaridade, país, estado, cidade de P e PC são iguais, e;

(2) Se (words ∩ concepts) ≥ (|concepts| / 2).

Em seguida, é feita uma análise no resultado da intersecção entre os dois vetores para verificar se o mesmo é maior ou igual ao tamanho do vetor *concepts*, pois provavelmente o vetor *words*, por conter os dados oriundos do perfil do usuário do SNS, terá mais componentes. Além disso, quando houver mais que 50%[8] dos componentes do vetor *concepts* no vetor *words*, implicará que o usuário usou boa parte dos conceitos extraídos da base cultural do OMCS-Br, e que por isso ele tem chances de estar falando sobre o assunto em questão.

O campo *words* de *P*, como dito anteriormente, representa as palavras que certo usuário de SNS usou em seu perfil. O seu preenchimento depende de qual SNS o usuário está inscrito. Caso seja o Orkut esse conjunto pode ser preenchido com as palavras que o usuário digitou no campo "Quem sou eu"[9]. No caso do Hi5 ele pode ser preenchido com o campo "Sobre mim"[10]. Enfim, depende do SNS que é aplicado a busca e da abordagem adotada na recuperação desses dados.

Quando *P* é considerado similar a *PC*, ele é adicionado a um conjunto $\beta = (P_1, P_2, ..., P_n)$ que representa todos os *Ps*, que representam usuários de SNS que têm chances de estar falando sobre o assunto, pois segundo a regra (1), definida anteriormente, pessoas que possuem o mesmo perfil podem utilizar o mesmo vocabulário quando se expressam sobre um mesmo assunto.

Na Figura 4.4 é apresentado o esquema de todo o método descrito. Nessa figura é possível que os dados do perfil dos usuários do Orkut são recuperados de várias áreas do perfil do mesmo.

[8] Esse valor foi definido em um primeiro momento apenas para verificar uma possível viabilidade do método, pois entende-se que um estudo maior deve ser feito para que esse número seja estipulado com maior precisão.
[9] Quem sou eu – esse campo é usado pelos usuários do Orkut para descrever algo sobre eles, como por exemplo, ideologia, descrição de hobbies, etc. Um exemplo desse campo é mostrado na Figura 2.3 área B.
[10] Sobre mim – esse campo possui as mesmas características que "Quem sou eu" do Orkut.

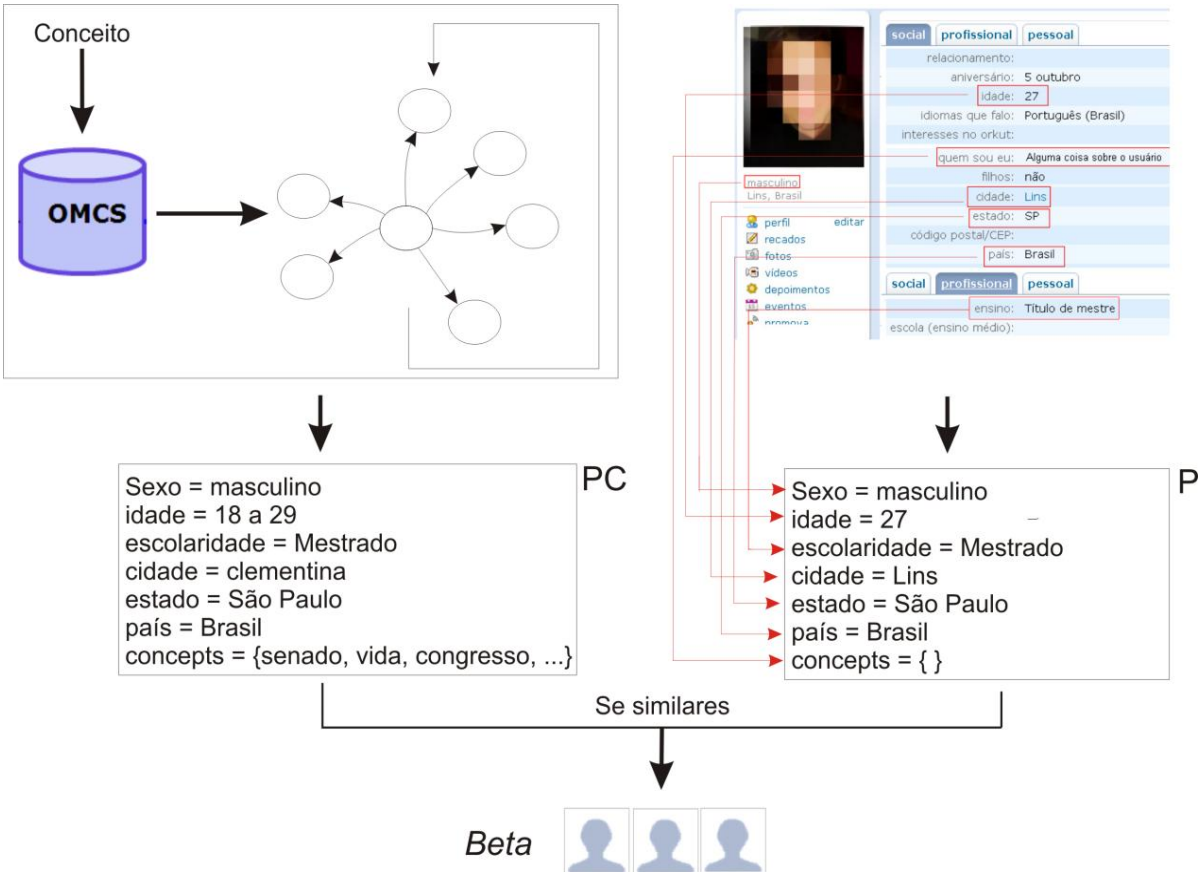


Figura 4. 4. Arquitetura do método quando considera o perfil e o vocabulário das pessoas.

Na próxima seção é apresentado um experimento com o objetivo de observar o funcionamento do método proposto. Esse experimento visa, não somente fazer uma contagem numérica dos dados, mas também analisar se as ocorrências dos conceitos (*concepts* $[w_1, ..., w_n]$) nos perfis (*words* $[w_1, ..., w_n]$), selecionados em β, reflete exatamente o assunto representado por *PC*.

4.2.3 Teste

O experimento foi realizado para observar o funcionamento do método e com isso também levantar subsídios que pudessem deixar mais claro o caminho a ser percorrido em direção a resolução do problema levantado por este trabalho. Por esse motivo que até esse momento não houve a participação de pessoas no experimento, pois o intuito é observar o funcionamento do método e os retornos obtidos.

Os passos adotados para o experimento se resumem em:

(1) Escolha de assuntos representativos;

(2) Recuperação dos (PCs) da base de conhecimento cultural do OMCS-Br;

(3) Escolha do SNS para a busca de perfis (P);

(4) Desenvolvimento de um aplicativo para a busca de P, e;

(5) Busca de pessoas que poderiam se interessar pelo assunto escolhido.

Os assuntos usados para a recuperação dos cincos *PCs* distintos foi escolhido aleatoriamente por um aluno de pós-graduação em ciência de computação da Universidade Federal de São Carlos. São eles: *festa, música, futebol, jogos e computador*. A Tabela 4.2 mostra os *PCs* conseguidos para cada um dos assuntos. Ressalta-se que o processo de obtenção de cada *PC* relacionado a cada assunto é feito um por vez.

Tabela 4. 2. Resultado da busca por PC para cada assunto.

Assunto	Idade	Sexo	País	Est.	Cid.	Escolaridade	Concepts
Festa	18 - 29	F	Brasil	SP	São Carlos	2º grau completo.	Bolo, amigo, música, comemoração, confraternização e aniversariante
Música	18 - 29	M	Brasil	SP	São Paulo	2º grau completo.	Ouvir, Cd, tocar, Discoteca, show de música, show musical, dançar, danceteria, casa.
Futebol	18 - 29	M	Brasil	SP	São Carlos	2º grau incompleto.	Jogar, Esporte, Estádio, Praticar, Assistir, torcer
Jogos	18 - 29	M	Brasil	SP	São Paulo	2º grau completo.	Assistir, fábrica de presente, ver, loja de presente, celular.
computador	18 - 29	M	Brasil	SP	São Carlos	2º grau completo.	laboratório de informática, escritório lan house, sala de informática, cyber café, trabalhar

O SNS escolhido foi o Orkut, pois a maior parte de seus usuários são brasileiros, como já dito na seção 2.3, e os colaboradores do projeto OMCS-Br, de onde são extraídos os *PCs*, também são brasileiros. Isso sugere que a expressão cultural encontrada na base do OMCS-Br pode ser também encontrada no perfil dos usuários do Orkut.

Para fazer a busca de perfis no Orkut foi desenvolvido um aplicativo que busca, por certo período de tempo, os perfis públicos (*P*) de usuários brasileiros e os armazena em um banco de dados. A infra-estrutura construída para essa busca se baseia em um analisador de HTML (*HyperText Markup Language*) que vasculha a estrutura da página do perfil do usuário, a fim de encontrar no campo "país" o dado "Brasil" e, quando esse é encontrado, o perfil é coletado.

Finalmente, ao término da coleta de dados têm-se cinco *PCs* (Tabela 4.2), sendo que cada um representa um assunto (*festa*, *música*, *futebol*, *jogos* e *computador*), além disso, há uma base de dados com uma grande coleção de *P*s extraídos do Orkut, que será utilizada para fazer as comparações com cada um dos *PCs*.

4.2.3.1 Resultados

O aplicativo, usado para buscar os perfis (*P*) no Orkut, trabalhou por um período de tempo de duas horas e trinta minutos. Ao final de sua execução foram obtidos aproximadamente 36.000 *Ps*. Nessa etapa foi considerada apenas a localização de *P*, isto é, foram buscados apenas os perfis dos usuários do Orkut onde o campo país estava preenchido com o dado "Brasil".

É válido mencionar que não foi usado individualmente cada *PC* para a busca dos perfis (*P*), pois o intuito inicial era buscar todos os perfis em que o "país" fosse "Brasil" e qualquer *PC* (Tabela 4.2) tem esse quesito. Por esse motivo considerou-se desnecessário a busca considerando todos os *PCs*, no entanto, cada um dos cincos *PCs* serão comparados com cada um dos 36.000 *Ps* recuperados.

Após a comparação, de cada um dos *PCs* com todos os 36.000 *Ps* em busca de similaridade, foi observado um resultado não satisfatório, pois houve de forma geral aproximadamente 0.019% de sucesso, isto é, conseguiu-se 7 usuários (*Ps*) com informações de sexo, idade, pais, estado, cidade, escolaridade iguais aos dos *PCs*. Além de encontrar igualdade com mais de 50% das palavras mantidas no campo do perfil "Quem sou eu" do Orkut. Ressalta-se que o (*P*) não precisava ser semelhante a todos os (*PCs*), se fosse semelhante a um *PC* era o suficiente, pois os assuntos escolhidos são tratados distintamente.

O primeiro *PC*, referente ao assunto *festa*, conseguiu identificar aproximadamente 0.0060% de *Ps*, entre os 36.000; o segundo, referente ao assunto *música*, 0.0030%; o terceiro, referente ao assunto *futebol*, 0.0060%; o quarto, referente ao assunto *jogos*, 0% e; o quinto, referente ao assunto *computador*, 0.0030% . Na Figura 4.5 são mostrados os resultados em um gráfico.

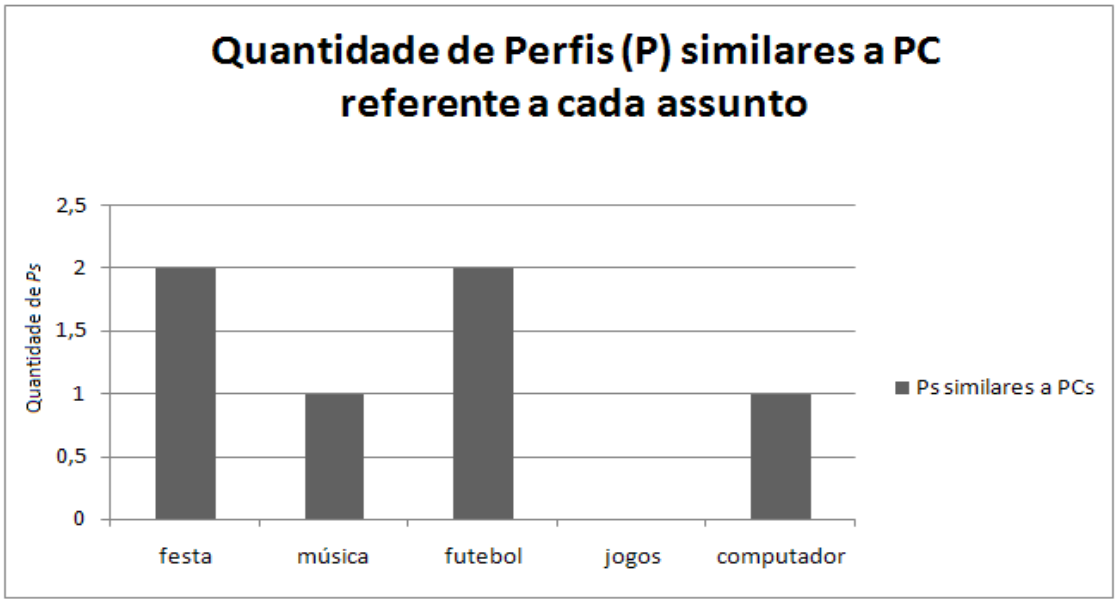


Figura 4. 5. Desempenho entre os assuntos usados no experimento a fim de identificar pessoas que possuem o mesmo perfil e compartilham o mesmo vocabulário.

Quando se trata apenas de comparação entre os dados do perfil (*idade, sexo, país, estado, cidade* e *escolaridade*) o resultado, de modo geral, é um pouco melhor em relação a quantidade 0.78% (281). O primeiro perfil de *PC*, referente ao assunto *festa*, conseguiu identificar aproximadamente 0.20% de *Ps* similares, entre os 36.000; o segundo, referente ao assunto *música*, 0.12%; o terceiro, referente ao assunto *futebol*, 0.20%; o quarto, referente ao assunto *jogos*, 0.12% e; o quinto, referente ao assunto *computador*, 0.12%. Na Figura 4.6 são mostrados os resultados.

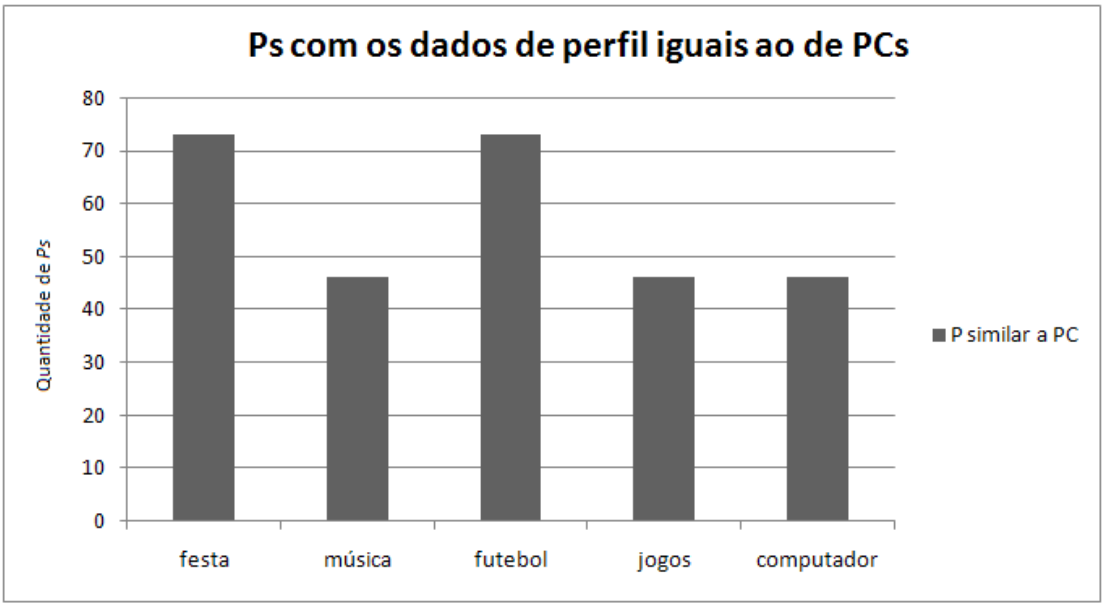


Figura 4. 6. Comparação entre os desempenhos dos assuntos considerando apenas o perfil dos usuários do

Na próxima subseção são analisados os dados para o levantamento de problemas e/ou subsídios para uma nova iteração.

4.2.4 Apontar falhas

O método apresentou alguns problemas no que diz respeito à utilização dos dados dos perfis dos colaboradores do projeto OMCS-Br. Por meio do teste descrito anteriormente foi possível observar que não se pode afirmar que dois usuários de certo SNS poderiam ter os mesmos interesses apenas pelo simples fato de possuírem a mesma *idade, localização, grau de escolaridade* e *sexo*.

Esse problema relatado foi percebido tanto no trabalho de Nunes (2008), quanto no teste realizado, pois os dados apresentados, na Figura 4.6, são indícios desse problema. Por exemplo, com relação ao assunto "jogos", na comparação apenas entre os dados do perfil foram identificados 0.12% (46) usuários, mas quando a comparação foi considerando o vocabulário dos usuários esse número caiu para 0% (Figura 4.5).

Dentre os campos do perfil, a *idade* foi o campo que ocasionou o maior problema, pois se considerar todos os usuários brasileiros do Orkut que estão inseridos dentro de uma das faixas etárias considerada pelo projeto OMCS-Br, entre 18 e 29 anos, o método deveria trabalhar com cerca de 18 milhões de usuários (34% de usuários brasileiro – fonte: www.orkut.com). Na expectativa que esses poderiam se interessar pelo assunto em questão devido à faixa etária, ou seja, algo inviável.

Outro problema percebido é sobre o uso do campo "quem sou eu" do perfil dos usuários do Orkut usado como fonte de dados. Nesse campo, como observado durante o experimento, as pessoas, muitas vezes, escrevem poesias e/ou letras de músicas. Esse tipo de comportamento pode ter prejudicado o desempenho do método, pois isso influencia na busca pelo real vocabulário do usuário.

Além disso, o campo *concepts* (oriundo da base do OMCS-Br) pode não possuir um relacionamento semântico coerente entre suas palavras, pois a forma como elas são extraídas da base do projeto OMCS-Br, por meio da ocorrência e não por relacionamento semântico direto uma com a outra, pode prejudicar a relação entre elas. Por exemplo, os conceitos "discoteca" e "casa", que representam o assunto "Música", não possuem relacionamento semântico aparente, além disso, não estavam relacionados um com o outro na base do OMCS-Br quando foram recuperados pelo método. Outro exemplo está relacionado ao assunto "jogos", também se percebe que algumas de suas palavras não possuem relacionamento semântico aparente. Nesse caso, mesmo reunindo um vocabulário sobre o

assunto é repetido o problema apresentado pelo trabalho de Chen (2009), discutido anteriormente na seção 4.2.2.

A forma como os usuários do Orkut são buscados pelo o método, usando apenas a localização (país = “Brasil”), demonstrou ser uma questão de sorte ao invés de uma busca coerente que usa parâmetros significativos. Considerando que o Orkut possui aproximadamente 60 milhões de usuários brasileiros, durante o processo podem vir vários perfis com o campo “quem sou eu” bastante significativo, ou vários insignificantes, pois não há um controle sobre isso. Isso indica que devem ser consideradas heurísticas mais específicas para a busca das amostras, como por exemplo, considerar, além do campo (país= “Brasil”), pelo menos um ou dois conceitos do vocabulário que representa o assunto para uma busca no campo “quem sou eu” do Orkut.

A partir desses comentários e conclusões observou-se a necessidade da evolução do método, ou seja, de uma nova iteração, que é discutida na próxima seção.

4.3 Iteração 2 – Identificando pessoas com um vocabulário comum

Considerando os problemas citados, a segunda iteração não faz uso dos perfis dos colaboradores do projeto OMCS-Br, no entanto, essa iteração tem como objetivo conseguir melhor relacionamento semântico entre as palavras que representa o vocabulário dos mesmos. Para tentar atingir esse objetivo, foi modificada a forma como o vocabulário é extraído da base do OMCS-Br, pois nessa iteração são consideradas apenas algumas relações de Minsky (ConceptuallyRelatedTo, ThematicKline e SuperThematicKline), ao invés de todas como na iteração anterior.

A mineração dos dados no Orkut, que antes era feita na página de perfil dos usuários, nessa iteração é feita considerando as comunidades. Dessa forma, se pretende evitar que os usuários omitam seus reais vocabulários como acontece no campo “quem sou eu” da página de perfil, como por exemplo, escrevendo letras de músicas e poesias. Nas comunidades, possivelmente os usuários expressam as suas opiniões, sendo assim, naturalmente escrevem considerando seu conhecimento, vocabulário, etc. Devido à utilização de comunidades, a seção 4.3.1 apresenta uma breve explicação sobre as comunidades do Orkut.

É válido mencionar que as modificações feitas no método, discutidas brevemente acima, são apresentadas em detalhes na seção 4.3.3.

4.3.1 Comunidades do Orkut

Segundo RHEINGOLD (1994) as comunidades virtuais são agregadores sociais que surgem das redes (Internet), para levar adiante discussões, sobre qualquer assunto, tornando-as públicas por certo período de tempo. Esse conceito é um fato nas comunidades do Orkut, pois os usuários discutem, através de fóruns, enquetes e troca de mensagens sobre os mais variados assuntos, como por exemplo, política, religião, música, entre outros.

Percebe-se também que alguns assuntos discutidos nas comunidades as fezes derivam de notícias de grande repercussão e importância, como por exemplo, o terremoto no Chile em 2010, ou notícias de corrupção no governo, etc.

Diferentemente de como os usuários usam seus perfis, nas comunidades, mais precisamente nos fóruns de discussão, eles expressam seus vocabulários e opiniões mais livremente e dinamicamente, pois, como pode ser observada na Figura 4.7, a forma de interação naturalmente proporciona aos usuários esse tipo de comportamento. Esse fato, como discutido anteriormente, geralmente não ocorre no perfil do usuário, mais precisamente na parte "quem sou eu", pois, muitas vezes, as pessoas apenas inserem um texto que se identificam, sem a necessidade de expressar seu vocabulário, cultura, etc.

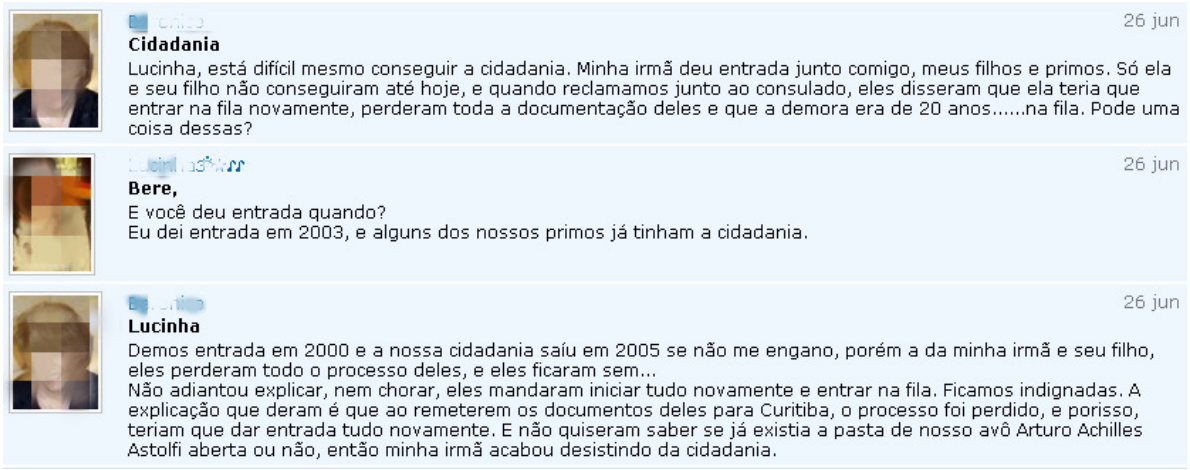


Figura 4. 7. Exemplo de postagem nas comunidades do Orkut.

Como pode ser observado na Figura 4.7, as discussões nas comunidades estão relacionadas com algum tipo de assunto, que frequentemente está associado a algum contexto. Devido a essa característica, foi percebido um provável potencial das comunidades do Orkut como fonte de vocabulário de usuários de SNSs, por isso, ao invés de minerar os dados no perfil dos usuários, surgiu a ideia de minerar os dados nos fóruns de discussões das comunidades. Esse processo é mostrado na seção 4.3.3.

4.3.2 Exposição do problema

Como descrito na iteração anterior, o método inicialmente foi inspirado em trabalhos que exploram o perfil e o vocabulário das pessoas a fim de definir um contexto que possa inferir similaridade entre elas. No entanto, o uso de perfis não obteve resultados satisfatórios, pois observou-se que não foi uma estratégia adequada considerar que pessoas que possuem o mesmo perfil possivelmente têm o mesmo vocabulário quando se trata de certo assunto.

Com o intuito de tentar resolver esse problema foi cogitada a possibilidade de considerar mais dados do perfil, como por exemplo, profissão, etc., porém, pela experiência com a iteração anterior, observou-se que aumentar a quantidade de dados do perfil não surtiria o efeito esperado, uma vez que não há possibilidade de garantir que as pessoas identificadas teriam o mesmo vocabulário e interesses em relação a certo assunto. Além disso, quanto mais dados fossem utilizados do perfil, mais complexa seria busca.

Em relação a explorar o vocabulário das pessoas, os problemas que mais se destacaram foram: usar as páginas do perfil dos usuários do Orkut como fonte de dados. Devido às pessoas não usarem o campo "Quem sou eu" como o esperado; utilizar as palavras, extraídas da base do OMCS-Br, considerando todas as relações de Minsky para representar o vocabulário, pois se percebeu que em alguns dos casos essas palavras apresentaram baixo relacionamento semântico aparente, o que ocasionou falta de representatividade do assunto em questão.

Apesar das falhas em relação ao uso do vocabulário, o maior problema foi percebido no uso do perfil, por isso, para essa iteração da evolução do método é desconsiderado o uso do perfil dos colaboradores do projeto OMCS-Br, passando apenas a explorar o vocabulário dos mesmos, sendo assim, o *PC* passa ser composto somente pelo vetor *concepts*.

Para essa iteração assumiu-se a seguinte regra:

(2) ***pessoas que falam sobre o mesmo assunto podem compartilhar o mesmo vocabulário.***

É válido reforçar que, como dito anteriormente na regra (1), as regras vão ser evoluídas à medida que problemas vão sendo identificados.

A partir dessa nova regra, o objetivo foi explorar uma forma de tentar resolver o problema do baixo acoplamento das palavras mantidas por *concepts*, que representam o vocabulário extraído da base do OMCS-Br, com o intuito de identificar pessoas em um contexto onde se possa perceber que elas estão falando sobre o mesmo assunto.

4.3.3 Resolução do problema

O trabalho de Granada *et al.* (2006) usa um método parecido ao de Chen (2009), pois sua proposta concentra-se em formar grupos de profissionais que sejam capazes de realizar uma determinada tarefa, a partir de uma análise no currículo Lattes (http://lattes.cnpq.br/) das pessoas, onde extrai as palavras mais usadas por cada uma.

Como no trabalho do Chen (2009), há um "saco de palavras" (vetor) para cada indivíduo e, tais palavras caracterizam a especialidade de cada pessoa. Nesse contexto, os indivíduos que possuem similaridade entre seus vetores podem ser considerados especialistas sobre o mesmo assunto.

Comparando este trabalho com o trabalho de Granada *et al.* (2006) é possível perceber uma certa diferença, pois aqui o objetivo é procurar pessoas que estão falando sobre o mesmo assunto e, não para identificar pessoas semelhantes para desenvolverem uma certa tarefa, contudo, o funcionamento de seu método é interessante para este trabalho, porque sua técnica pode ser inspiradora para identificar pessoas que podem compartilhar um vocabulário comum em SNSs.

Uma observação importante sobre o trabalho de Granada *et al.* (2006) é que ele conseguiu fazer com que as palavras do "saco de palavras" de cada indivíduo tenham relacionamento implícito[11] entre elas. Pois, ao fazer a busca no currículo Lattes de uma pessoa, provavelmente encontrará palavras referente a uma linha de pesquisa, caso ela seja um pesquisador ou um especialista, etc.

Essa contextualização, mesmo que de forma implícita, conseguida por Granada *et al.* (2006) é um indício de que é possível resolver o problema de falta de relacionamento semântico, percebida no experimento da versão anterior do método, entre as palavras do vetor *concepts* de *PC*.

[11] Relacionamento implícito entre palavras refere-se a um relacionamento que ocorre entre um conjunto de palavras pelo fato delas pertencerem a um mesmo contexto. Por exemplo, as palavras software, Java, projeto e usabilidade, possuem um relacionamento implícito devido eles serem usadas no contexto da informática.

Quando há uma discussão sobre certo assunto, independente de qual seja, foi observado na pesquisa feita por este trabalho nos Sites de Redes Sociais, que as pessoas usam um conjunto específico de palavras para desenvolvê-lo. As principais palavras usadas, de alguma forma, são conectadas semanticamente. O trecho de texto extraído de uma postagem de um usuário do Facebook justifica essa observação:

"Nesta semana temos Dark Void, que conta a história do piloto de avião de carga que caiu no misterioso triângulo das Bermudas."

Nesse caso, se considerar apenas os substantivos do texto (*Semana, Dark Void, história, piloto, avião, carga, triângulo, Bermudas*) é possível observar que o assunto trata-se do lançamento de um filme ou de um livro que conta certa história e, que os substantivos possuem algum tipo de relacionamento semântico implícito que pode fornecer essa informação. Observa-se que os cinco últimos substantivos da listagem podem caracterizar bem o relacionamento entre as palavras.

Esse tipo de relacionamento semântico apenas entre os substantivos, que se sabe que há, mas não se sabe precisar como (HAVASI, 1997), é previsto em uma das classes de relações de Minsky, as chamadas *K-lines*. As relações que fazem parte dessa classe, bem como suas definições são apresentadas na Tabela 4.3.

Tabela 4. 3. Definição das relações de Minsky da classe das K-lines.

Relação	Definição
ConceptuallyRelatedTo	E um tipo de relação que diz que existe uma relação entre os dois conceitos, mas não é possível determinar qual é (HAVASI, 1997), ou seja, eles são relacionados, mas por um caminho desconhecido. Exemplo: *ConceptuallyRelatedTo (história, avião)*. Apenas quem leu a frase de onde eles foram extraídos sabe dizer o tipo de relacionamento.
ThematicKline	Define um relacionamento entre coisas sobre o mesmo tema, ou seja, alguma coisa que lembra outra (LIU, 2004). Exemplo: *ThematicKline (avião, Bermudas).*
SuperThematicKline	Unifica o tema com suas variações (LIU, 2004). Exemplo: "Lançamento" é um super tema para "Lançamento de filme" e "Lançamento de avião", *SuperThematicKline (Lançamento, filme)* e *SuperThematicKline (Lançamento, avião).*

Na Tabela 4.4, há uma cópia de uma das linhas da Tabela 4.2 com o vetor c*oncepts* conseguido da base de conhecimento do OMCS-Br. Provavelmente suas palavras não estavam relacionadas diretamente umas as outras através de alguma das relações de Minsky.

Tabela 4. 4. Exemplo de PC. Uma reprodução de uma linha da Tabela 4.2.

Idade	Sexo	País	Estado	Cidade	Escolaridade	Conceitos
18 - 29	M	Brasil	SP	São Paulo	Segundo grau completo.	Bolo, amigo, música, comemoração, confraternização e aniversariante.

Na Figura 4.8 é mostrado esse fato, pois é possível observar que entre "*bolo*" e "*comemoração"* há um caminho semântico que implicitamente pode ser resumido por uma relação da classe das *K-lines - ThematicKline (bolo, comemoração)*.

Esse fato também é percebido no conjunto de palavras que as pessoas usam em suas conversas. Por exemplo, no trecho de texto exemplificado anteriormente os substantivos "*avião*" e "*Bermudas*" possuem um relacionamento semântico implícito, assim como "*bolo"* e "*comemoração"* e, por isso, podem ser mapeados pela relação *ThematicKline (avião, Bermudas)* .

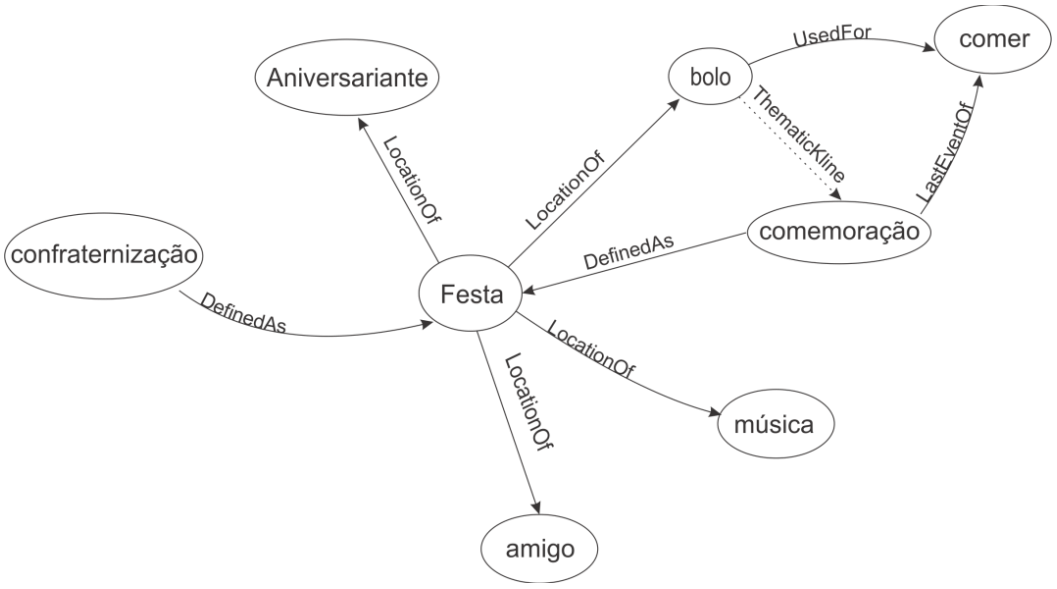


Figura 4. 8. Exemplo de uma associação implícita entre conceitos.

Considerando esse fato, uma alternativa para tentar resolver o problema relacionado ao baixo acoplamento das palavras de *concepts*, pode ser o uso somente da classe das *K-lines* no processo de busca por vocabulário na base do OMCS-Br, Dessa forma, um melhor relacionamento semântico entre as palavras que representam o vocabulário e consequentemente o assunto, podem ser conseguido, fazendo com que todo o conteúdo de *concepts* esteja relacionado semanticamente.

Para que isso seja feito inicialmente o método recebe como entrada um conjunto de palavras que representa um assunto. Para que seja garantido algum

relacionamento semântico entre elas, essas palavras são retiradas de uma ou mais sentenças, sobre um mesmo assunto. Diferentemente da iteração anterior que o assunto era representado apenas por uma palavra.

Esse conjunto de palavras é parte do esforço para melhorar o relacionamento semântico entre as palavras que serão mantidas por *concepts*, pois, ao usar palavras que já têm um relacionamento prévio para minerar conhecimento cultural da base do OMCS-Br, pode-se ter um acoplamento maior entre os conceitos recuperados da base, e consequentemente uma maior representatividade sobre o assunto.

Para exemplificar a obtenção do conjunto de palavras considere as seguintes sentenças: *"Lula defende Sarney e diz que denúncias não têm fim"*[12] e *"Lula critica seqüência de denúncias sobre o Senado e defende Sarney"*[13]. O conjunto de palavras, que a partir desse ponto será chamado de *Subs*, é formado pelos substantivos ignorando as repetições retiradas das sentenças. Nesse exemplo *Subs = {Lula, Sarney, denúncias, fim, sequencia, Senado}.*

Para formar *Subs* cada uma das sentenças é submetida ao Curupira (Martins et. al, 2003), que indica qual a categoria que cada palavra está incluída. Nesse caso, são consideradas apenas as palavras categorizadas como substantivos (`SUBST` – etiqueta usada pelo Curupira para marca uma palavra como substantivo).

Nesse ponto é interessante retomar o assunto sobre semântica, discutido anteriormente, para chamar atenção sobre as palavras mantidas por *Subs*. Percebe-se que elas, mesmo não estando mais ligadas por meio de verbos, artigos, preposições, etc., conseguem refletir sobre um determinado assunto, que nesse caso é política, além disso, é possível mapear, usando permutação, cada uma delas para alguma relação da classe das *K-lines*. Por exemplo, *ThematicKline (Lula, Sarney)*, sabe-se que há uma relação, mas sem os componentes retirados da sentença não é possível precisar.

Dessa forma, há indícios de que o primeiro objetivo pode ser alcançado, isto é, pelo menos as palavras, que serão usadas para coletar conceitos relacionados da base do OMCS-Br, estão de alguma forma relacionadas entre si.

A segunda etapa do método usa cada uma das palavras de *Subs* para extrair conceitos relacionados da base cultural do projeto OMCS-Br.

---

[12] Manchete de uma notícia publicada pela Folha On Line no dia 17 de junho de 2009.
[13] Manchete de uma notícia publicada pela Gazeta do Povo On Line no dia 17 de junho de 2009.

Esse procedimento objetiva buscar informações apenas das relações da classe das *K-lines*, isto é, a busca é fixa com cada uma das palavras e cada uma das três relações. Por exemplo, uma busca com a palavra "*Lula*" $\in$ *Subs* é tratada como é mostrado na Tabela 4.5. A primeira coluna apresenta como a busca é executada e a segunda, o resultado.

Tabela 4. 5. Exemplo de uma busca por conceitos na base do OMCS -Br usando somente a classe das K-lines.

Parâmetro	Retorno em X ou Y
ThematicKline (Lula, Y)	Presidente
ThematicKline (X, Lula)	Votação, político
ConceptuallyRelatedTo (Lula, Y)	Não houve retorno
ConceptuallyRelatedTo (X, Lula)	voto
SuperThematicKline (Lula, Y)	Não houve retorno
SuperThematicKline (X, Lula)	Não houve retorno

Todos os conceitos conseguidos com a busca (segunda coluna da Tabela 4.5) são agrupados em um conjunto, no qual é chamado de *Conc*. Por fim, o vetor de *concepts* que se pretende evoluir é construído da seguinte forma: *concepts* = *Subs*$\cup$*Conc*.

O objetivo de construir um vetor de palavras conectadas semanticamente é conseguido com a obtenção de *concepts*. Considera-se que as palavras que pertencem a *concepts* estão conectadas semanticamente, porque parte delas foram extraídas de sentenças onde já estavam conectadas e, a busca pela base de conhecimento cultural usando a classe das *K-line*, adicionou outras palavras, que de acordo com o conhecimento cultural dos colaboradores do projeto, estão relacionadas a elas.

Após esse processo, espera-se ter um conjunto de palavras que estão acopladas, ou seja, conectadas semanticamente. Essas palavras serão usadas como parâmetro a fim de identificar pessoas, em SNSs, que as usam em seus vocabulários, ou seja, que estão falando sobre o assunto em questão.

A busca é realizada nas postagens dos usuários nas Comunidades do Orkut, com o objetivo de encontrar um percentual elevado das palavras mantidas por *concepts*, em uma única postagem de um usuário.

Considerando que as palavras de *concepts* estão conectadas semanticamente e, por isso, podem refletir sobre certo assunto, ao encontrar uma postagem de um usuário que contém, em seu conteúdo, boa parte das palavras mantidas por *concepts*, pode-se considerar que esse usuário está falando sobre o assunto no qual *concepts* é capaz de representar.

Quando são encontradas postagens que possuem palavras similares às existentes em *concepts*, o endereço do perfil do usuário, responsável pela postagem, é selecionando e adicionado ao conjunto $\beta = (P_1, P_2, ..., P_n)$. Cada P, dessa vez, representa o endereço de um usuário que possivelmente tem chances de estar falando sobre o assunto representado pelas palavras mantidas por *concepts*.

A medida de similaridade é conseguida por meio da medida denominada Coeficiente de Jaccard (Figura 4.9). Esse coeficiente mede a similaridade entre dois conjuntos, que nesse caso são palavras. Considera-se que dois conjuntos são similares apenas quando a medida ultrapassa o valor de 0.5, ou seja, 50%. A seguir, cada variável considerada pelo coeficiente de Jaccard é descrita:

- *C*: número de palavras comuns entre os dois conjuntos;
- *Concepts*: número de palavras do primeiro conjunto;
- *Post*: número de palavras do segundo conjunto e;
- *Cj*: o quanto os dois conjuntos são similares.

$$Cj = \frac{C}{Post + Concepts - C}$$

Figura 4. 9. Coeficiente de Jaccard.

Nesse caso, é medido o quanto o vetor *concepts* é similar ao vetor *Post* que é um vetor conseguido a partir das palavras extraídas das postagens dos usuários. Com o intuito de observar essa similaridade, foram buscados pares de palavras, no conteúdo das postagens, que fazem parte do vetor *concepts*.

Essa heurística tenta resolver o problema relacionado à busca das amostras da iteração anterior. Pois, dessa forma são recuperadas apenas as postagens que estão relacionadas ao assunto mantido por *concepts*, e não qualquer uma, como era feito na versão anterior do método.

Na Figura 4.10, há um exemplo de uma postagem recuperada pela busca. Percebe-se que todas as palavras em destaque fazem parte do vetor *concepts*, mas os pares usados para a busca foram as palavras “Sarney” e “Senado”.

A rendição do presidente se deu naquela célebre entrevista concedida em Paris, em 2005, nos tempos em que a corrupção causava ainda algum constrangimento. Sem os corretivos vindos de cima, a turma do baixo, do médio e do alto clero da base aliada sentiu-se mais livre do que nunca. Sempre que um de seus integrantes está prestes a se afogar, eis que surge o presidente, solidário, oferecendo o conforto de suas palavras amigas. Nem precisa ser compadre de pitar cigarrilha, como o leal companheiro Delúbio Soares, estrela do mensalão. Pode ser do PMDB, do PP ou do PTB. Pode até ser, vá lá, um "grande ladrão", adjetivo com o qual Lula descrevia o senador José Sarney quando este era presidente da República. Há cinco meses o Congresso Nacional enfrenta uma infindável onda de escândalos. Ela envolve parlamentares e altos funcionários com mordomias, nepotismo e suspeitas de corrupção. Aos 79 anos de idade, 54 de política, Sarney, o mais longevo e experiente dos políticos brasileiros, é apontado como mentor e beneficiário da máquina clandestina que operava a burocracia do Senado. Inerte diante das denúncias, o senador tentou defender-se no plenário, com argumentos tão frágeis quanto os azulejos portugueses de São Luís.

Figura 4. 10. Exemplo de uma postagem recuperada por meio de buscas por pares de palavras.

À medida que são encontrados os pares de palavras, pertencentes ao vetor *concepts*, nas postagens, eles são selecionados e passam por um novo processo que visa verificar quais as outras palavras de *concepts* também ocorrem.

Cada uma das palavras encontradas, com exceção das repetições, é armazenada no vetor *Post*. Por exemplo, nesse caso *concepts = {Lula, Sarney, denúncias, fim, sequencia, Senado, ladrão, senador, congresso, deputado}* e *Post = {Lula, Sarney, denúncias, Senado, ladrão, senador, congresso}* (Figura 4.10). Ressalta-se que o vetor *Post* é dinâmico, pois a cada postagem nova que é conseguida, suas palavras são alteradas.

Para identificar a similaridade entre esses conjuntos existe uma sequência no método. Primeiro, são buscadas as postagens usando os pares de palavras; segundo, é construído o vetor *Post*; terceiro, é medida a similaridade entre *Post* e *concepts* e, finalmente, caso a similaridade seja maior que 50%, é recuperado o endereço do perfil do usuário e adicionado ao conjunto *β*. Esse processo é repetido novamente fazendo com que o vetor *Post* obtenha novas palavras vindas de uma nova postagem (veja Figura 4.11).

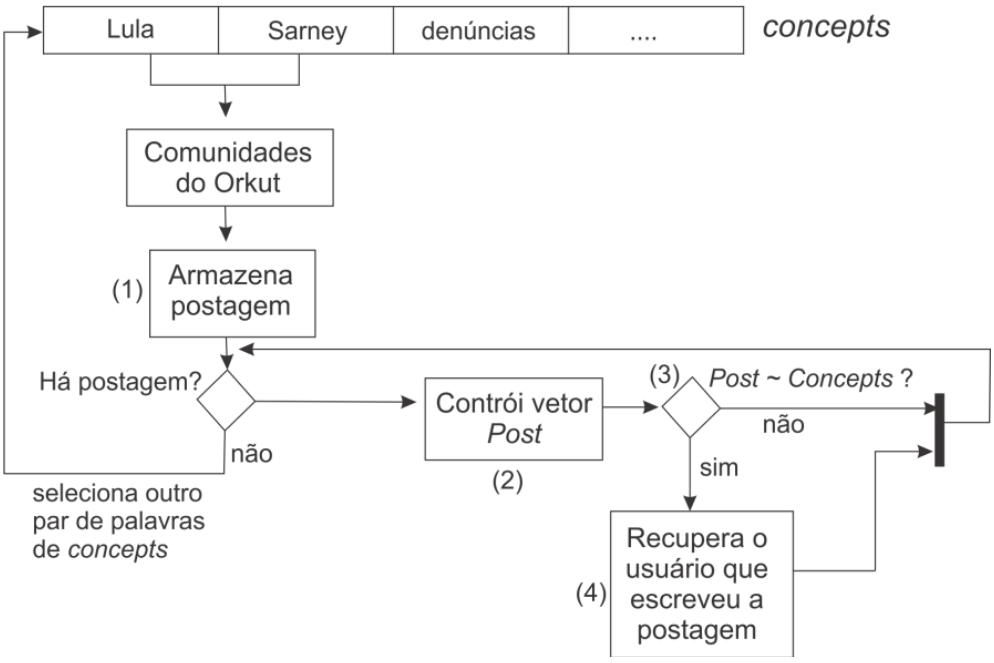


Figura 4. 11. Arquitetura da busca do método, quando considera apenas o vocabulário das pessoas.

A execução do método termina com um conjunto de usuários do Orkut, representados por cada $Ps \in \beta$, dos quais possivelmente estavam falando sobre o assunto usado na busca, pois quando se expressam em relação a tal, utilizam um conjunto de palavras, que conectadas semanticamente, possivelmente representam o assunto de maneira expressiva.

Como pode ser observada esta nova versão do método conseguiu fazer com que as palavras que representam o vocabulário das pessoas fossem melhor relacionadas semanticamente, como fez de forma implícita Granada *et. al* (2006), discutido anteriormente.

A diferença das duas iterações é que o "saco de palavras" (vetor *concepts*) é construído a partir de uma ou mais sentenças e enriquecido com o conhecimento cultural relacionado. Com essa estratégia, espera-se representar um assunto de forma mais coerente através do relacionamento semântico entre as palavras, algo que Granada *et. al* (2006) conseguiu apenas implicitamente.

4.3.4 Teste

Foi conduzido um experimento para verificar se houve melhora nos resultados do método em relação à versão apresentada na instância da iteração anterior da metodologia de trabalho. Os passos adotados para isso se resumem em:

(1) Escolha de sentenças que representa um assunto;
(2) Construção de concepts com suas palavras semanticamente relacionadas;
(3) Desenvolvimento de um aplicativo que busca postagens no Orkut e as compara com o concepts em busca de similaridade;

As sentenças usadas como sementes para a construção de *concepts* foram extraídas das manchetes de três portais de notícias on-line do Brasil: Gazeta do Povo on-line (www.gazetadopovo.com.br), Folha on-line (www.folha.com.br) e G1 (www.g1.com). Essa abordagem, de escolher manchetes de portais de notícias, foi adotada para evitar qualquer tipo de influência em relação à construção de *concepts*.

Os portais proveram as seguintes sentenças respectivamente: *"Lula critica sequência de denúncias sobre o Senado e defende Sarney", "Lula defende Sarney e diz que denúncias não têm fim"* e *"Lula pede apuração correta e tratamento diferenciado a Sarney"*. A partir dessas sentenças, cria-se o conjunto de palavras *Subs = {Lula, Sarney, denúncias, fim, sequencia, senado, apuração, tratamento}* que depois de submetido à base de conhecimento cultural conseguiu recuperar *Cons = {político, votação, presidente, voto,*

decisão, deputado, câmara, desonesto, senador, ladrão}, que unido com *Subs* produz *concepts = {Lula, Sarney, denúncias, fim, sequencia, senado, apuração, tratamento, político, votação, presidente, voto, decisão, deputado, câmara, desonesto, senador, ladrão}.*

Para usar *concepts* como parâmetro para a busca de postagens foi usada basicamente a mesma estratégia da iteração anterior, ou seja, foi desenvolvido um aplicativo que busca por certo período de tempo, as postagens que possuem pelo menos duas palavras de *concepts*. Buscaram-se pelo menos duas palavras por postagem para evitar que sejam recuperadas postagens com pouco significado em relação ao assunto procurado. Depois dessa pré-busca é que se analisaram as postagens recuperadas em busca de pelo menos 50% das palavras mantidas por *concepts* no conteúdo de cada uma.

A infra-estrutura construída para isso, também como na iteração anterior, se baseia em um analisador de HTML que vasculha a estrutura das páginas dos fóruns de discussão das comunidades do Orkut. O resultado da medida de similaridade entre *concepts* e as palavras extraídas das postagens selecionadas no processo de busca são descritas na próxima seção.

4.3.4.1 Resultados

Como na iteração anterior o aplicativo foi executado por um período de tempo, dessa vez por uma hora e meia. Ao final de sua execução conseguiu-se aproximadamente 21.900 comparações.

Após as comparações, foi possível observar que houve 0.073% de sucesso, ou seja, o método conseguiu identificar 16 postagens similares ao vetor *concepts* entre as aproximadamente 21.900 comparadas. Isso quer dizer que foi possível identificar 16 usuários do Orkut que usaram em algum momento, em suas discussões nas comunidades, 50% ou mais das palavras mantidas por *concepts*. A distribuição entre a quantidade de postagens identificadas e o número de palavras de *concepts* existente em cada uma delas pode ser visualizada na Tabela 4.6.

Tabela 4. 6. Quantidade de postagens identificadas e o número de palavras de concepts existente em cada uma delas.

Grupo	Número de postagens	Número de palavras de *concepts*	(%)
Grupo 1	9	9	50
Grupo 2	4	10	55.55
Grupo 3	2	11	61.11
Grupo 4	1	12	66.66

O primeiro grupo é composto por 9 postagens e em cada uma delas foram identificadas 9 (50%) palavras de *concepts,* no qual é composto por 18 conceitos. No segundo grupo há 4 com 10 (55.55%) palavras de *concepts*; no terceiro, 2 com 11 (61.11%) palavras e; finalmente no quarto, 1 com 12 (66.66%) palavras.

Entre as postagens desprezadas, ou seja, as que não possuíam pelo menos 50% das palavras de *concepts*, houve um grupo interessante de 32 postagens com 8 (44.44%) palavras de *concepts*.

4.3.5 Apontar falhas

Em relação à quantidade, essa iteração teve resultados melhores do que a anterior, pois nessa foram identificados dezesseis usuários para um assunto. Na anterior, o máximo que conseguiu foi identificar dois usuários por assunto (festa e futebol), no entanto, é preciso esclarecer que esse resultado pode depender do assunto em questão, pois talvez com outro essa versão não conseguisse identificar usuário algum.

Uma falha identificada, nessa versão do método, está relacionada com o falso-positivo que pode ocorrer, pois as palavras que pertencem a *concepts* podem ser encontradas na íntegra em certa postagem, mas em contextos ou assuntos completamente diferentes.

Por exemplo, a sentença *"Lula critica seqüência de denúncias sobre o Senado e defende Sarney"* usada como semente para a geração de *concepts* possui os seguintes substantivos: [Lula, sequencia, denúncias, senado, Sarney]. Em uma outra sentença *"Lula elogia sequencias de denúncias que Sarney fez no Senado"*, que pode ser encontrada em uma postagem, produz o mesmo conjunto de substantivos, mas está com um outro sentido.

Essa falha existe por que apesar do apelo semântico fornecido pela classe das *K-lines* para construir *concepts*, não se pode garantir que as palavras mantidas por ele encontradas nas postagens, quando conectadas através de verbos, preposições e outros componentes da oração, estão relacionadas com o assunto das sentenças usadas como sementes para gerá-lo. Isso é explicado pela diversidade de assuntos discutidos nas comunidades do Orkut, pois pode acontecer das mesmas palavras serem usadas para expressar um determinado assunto em diferentes perspectivas, bem como expressar assuntos diferentes.

Esse tipo de problema certamente não foi percebido por Granada *et. al* (2006), trabalho que influenciou o desenvolvimento da versão atual do método, porque a plataforma usada por ele na mineração dos dados (Lattes) é um ambiente direcionado a um grupo

especifico de pessoas, ou seja, pesquisadores. Público diferente do Orkut, que possui usuários mais heterogêneos.

Isso leva a conclusão que, mesmo com o apoio semântico fornecido pelas *K-lines* e das sentenças usadas como sementes, deve-se evoluir o método para que se consiga expressar ainda melhor a semântica entre as palavras de *concepts*, para poder afirmar com melhor exatidão que os usuários de SNSs selecionados, como o uso de *concepts*, estão falando sobre o mesmo assunto.

4.4 Iteração 3 – Identificando pessoas que falam sobre o mesmo assunto

Esta iteração se concentra em evoluir a abordagem usada na iteração anterior. Para isso, nesse ponto do trabalho, é inserido o conceito de meta-relação semântica, que será explicado posteriormente, para tentar resolver o problema de falso-positivo durante as buscas em SNSs.

Para um melhor desempenho do método, o *parser* Curupira foi substituído[14] pelo *parser* PALAVRAS, devido a inconsistência do Curupira em alguns testes, no qual o PALAVRAS mostrou-se mais eficiente. O PALAVRAS será apresentado com maiores detalhes na seção 4.4.2.2.

Finalmente, a última grande alteração no método é a inclusão de uma base de sinônimos para melhorar a abrangência da busca por pessoas que estão falando sobre o mesmo assunto em SNSs. Os detalhes de todas as alterações são descritos nas próximas seções.

4.4.1 Exposição do problema

A versão atual do método, evoluída a partir da primeira iteração, define um vetor de palavras relacionadas semanticamente para usá-lo como parâmetro em busca de usuários em SNS que utilizam as mesmas palavras quando se expressam. Dessa forma, até os testes descritos anteriormente, acreditava-se que as pessoas encontradas poderiam estar falando sobre o mesmo assunto.

Após os testes, observou-se de modo geral um resultado satisfatório, pois houve algumas melhoras do método descrito da segunda iteração em comparação com o método da primeira iteração. Porém, algumas falhas foram identificadas, como por exemplo,

[14] A substituição do parser Curupira pelo parser PALAVRAS também foi influenciada pelos testes apresentados em seminários pela Profa. Dra. Helena de Medeiros Caseli, especialista em Linguística Computacional, do Departamento de Computação da UFSCar.

falso-positivo, comentado na subseção 4.3.5 usada para apontar falhas. E, não foi possível afirmar com precisão que as pessoas selecionadas pelo método estavam falando sobre o assunto em questão.

O desafio dessa iteração foi resolver o problema relacionado ao falso-positivo e também afirmar com uma melhor precisão que as pessoas identificadas estão falando sobre o assunto em questão. Dessa forma, o objetivo foi tentar melhorar ainda mais o relacionamento semântico entre as palavras do vetor *concepts,* permitindo que elas tenham uma melhor representatividade sobre o assunto em questão.

4.4.2 Resolução do problema

As orações, exemplificadas no experimento anterior (subseção 4.3.4), que possibilitaram a identificação das falhas em relação ao falso-positivo, *"Lula critica seqüência de denúncias sobre o Senado e defende Sarney"* e *"Lula elogia sequências de denúncias que Sarney fez no Senado",* traz indícios que os verbos devem ser levados em consideração, pois dessa forma é possível fazer uma distinção entre as sentenças, algo que não seria tão simples de fazer apenas com os substantivos.

Por exemplo, uma maneira de considerar os verbos poderia ser por meio de relações semânticas, em que os mesmos seriam os agentes semânticos da relação, assim, as duas orações citadas acima poderiam gerar as seguintes relações semânticas:

- *critica (Lula, sequência denúncias senado) e defende (Lula, Sarney);*
- *elogia (Lula, sequências denúncias Sarney fez Senado).*

Observa-se que dessa forma, sem utilizar somente os substantivos, é possível identificar computacionalmente a diferença entre elas, basta iniciar a comparação pelos verbos, além disso, os substantivos não ficam desconectados, pois há o verbo para realizar a ligação, com o intuito de melhorar ainda mais o relacionamento entre eles.

Um trabalho que utiliza esse tipo de recurso é de Bollegala (2009). Ele propõe uma busca na web, no qual retorna a relação semântica entre as palavras-chaves usadas. Um dos exemplos que ele utiliza para ilustrar seu método é a possibilidade de uma busca com as palavras-chaves *Google* e *YouTube*. A relação semântica obtida é *acquirer (Google, YouTube*), da qual foi extraída de vários trechos de textos , como por exemplo: "*Google*

adquire Youtube por $1.65 bilhões em ações. Acordo criará novas oportunidades para usuários..."[15] (BOLLEGALA, 2009, tradução nossa).

Essa lógica descrita por Bollegala (2009) foi uma inspiração para tentar melhorar a semântica entre as palavras do vetor *concepts*, usado para identificar pessoas em SNS que estão falando sobre o mesmo assunto. A evolução do método se concentra em manter *concepts*, mas agora como um vetor de estruturas, onde cada uma representa uma relação semântica oriunda de uma oração escrita em linguagem natural, assim, PC = (*concepts* $[mr_1,..., mr_n]$) – onde mr_n é uma estrutura – junção de um verbo, substantivos e complementos.

Com essa estrutura, a busca passou de uma simples comparação entre palavras para uma comparação entre relações semânticas. Nesse caso, há a possibilidade de assumir a seguinte regra:

> ***(3) Se duas pessoas se expressam exatamente da mesma forma***[16] ***sobre um assunto, possivelmente elas estão falando sobre o assunto.***

Nesse contexto, é possível identificar pessoas que estão falando sobre o mesmo assunto, bem como considerar o consenso de cada uma em relação ao assunto. Por exemplo, quando se consegue identificar duas pessoas que se expressam de forma que suas sentenças gerem tal relação semântica: *critica (Lula, sequência denúncias senado), p*ode-se dizer que elas possuem o mesmo consenso e estão falando sobre o mesmo assunto.

Dessa forma, o objetivo dessa iteração é extrair relações semânticas das orações usadas como sementes para gerar *concepts* e, usar a base de conhecimento cultural do projeto OMCS-Br para extrair novas relações semânticas, a fim de enriquecer culturalmente a representação do assunto provido por *concepts*.

Para facilitar o entendimento sobre como fazer isso, na seção 4.4.2.1 é apresentado como é a relação semântica que essa iteração adota; na seção 4.4.2.2 é apresentado o parser PALAVRAS, uma das ferramentas usada na evolução do método; na seção 4.4.2.3 é mostrado como as relações semânticas são construídas com o uso do

---

[15] "*Google to acquire YouTube for $1.65 billion in stock. Combination will create new opportunities for user...*"

[16] Quando se usa a afirmação "expressam exatamente da mesma forma" para este trabalho significa que duas pessoas, quando falam de um mesmo assunto, usam as mesmas palavras principais para se expressarem. Por exemplo, se a Pessoa 1 diz: "*Lula critica seqüência de denúncias sobre o Senado e defende Sarney*", e a Pessoa 2 diz: "*Lula critica seqüência de denúncias no Senado além de defende Sarney*"; ambas estão se expressando exatamente da mesma forma. A mudança está nas preposições, advérbios e etc, mas não nos substantivos e verbos, que são considerados por este trabalho como as palavras principais de uma sentença ou oração.

PALAVRAS; na seção 4.4.2.4 é apresentado como a base de conhecimento cultural do OMCS-Br é usada para enriquecer culturalmente o assunto representado por *concepts,* bem como a utilização da base de sinônimos*;* e finalmente, na seção 4.4.2.5 é mostrado a evolução do método usando os conceitos apresentados nas seções 4.4.2.1 à 4.4.2.4.

4.4.2.1 Relações semânticas

As relações semânticas propostas por Bollegala (2009) são extraídas de trechos de textos em uma oração, no entanto, o método, proposto aqui, é estimulado por um conjunto de orações, pois há a necessidade de construir relações semânticas usando todas as principais palavras de uma oração, para que não se perca o assunto em questão.

A partir desse ponto todas as relações semânticas geradas serão chamadas de meta-relação – *mr*. Uma *mr* é capaz de representar a semântica de uma oração somente quando ela é composta por *sujeito, verbo* e *complemento.* Essa restrição é imposta porque uma *mr* é composta por *v, s* e *c,* ou seja, *mr = v (s, c),* onde:

- *v* representa o *verbo* da oração, a semântica de *mr;*
- *s* representa o *sujeito* da oração, o agente de uma ação, pois nas *mr* se considera apenas o aspecto semântico da oração*;*
- *c* representa o *complemento* da oração, o elemento determinado em relação à *s.*

Dessa forma, não há possibilidade de construir uma *mr* com a falta de um dos componentes da oração (*sujeito, verbo* e *complemento).* Para exemplificar, considere a oração *"Lula defende Sarney"*. Uma *mr* construída a partir dela tem a seguinte forma: *defender (Lula, Sarney)* (veja a representação gráfica na Figura 4.12(a)). Observe que a *mr* elimina as *stop words*[17] (com exceção da palavra "*não*" e verbos), que são palavras que não alteram o sentido e o significado da oração.

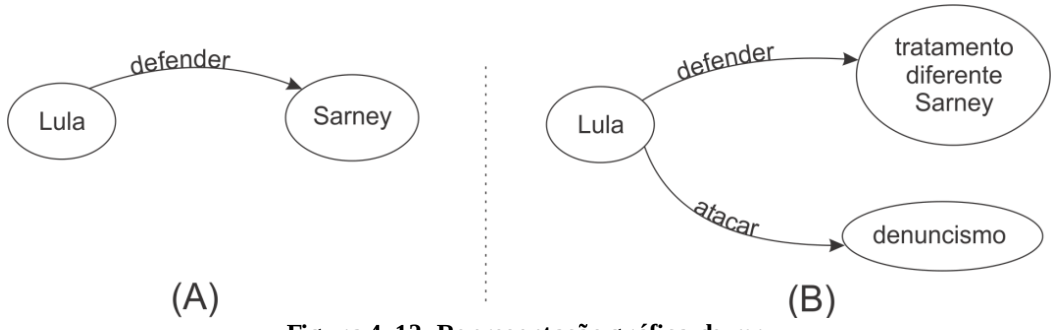


Figura 4. 12. Representação gráfica de *mr*.

[17] Alguns exemplos dessas palavras são artigos, preposições, etc., e podem ser encontradas em: http://snowball.tartarus.org/algorithms/portuguese/stop.txt.

Uma oração pode gerar uma ou mais *mr*, isso depende da forma e da carga semântica que ela possui. Por exemplo, *"Lula ataca denuncismo e defende tratamento diferente para Sarney"* gera duas *mr: atacar (Lula, denuncismo)* e *defender (Lula, tratamento diferente Sarney)* (veja Figura 4.12 (b)).

Isso é possível porque a oração em questão pode ser "dividida" em duas, ou seja, "*Lula ataca denuncismo*" e *"Lula defende tratamento diferente para Sarney",* então, qualquer oração que possa ser "dividida" em duas ou mais, é uma potencial geradora de duas ou mais *mrs*. A Tabela 4.7 mostra outros exemplos de geração de *mr* a partir de diferentes tipos de oração.

Tabela 4. 7. Exemplos de *mr* geradas a partir de orações.

Oração	***mr***
"Lula defende Sarney"	*defender (Lula, Sarney)*
"Lula ataca críticas sobre denúncias no Senado e protege corruptos"	*atacar (Lula, critica Senado)* *proteger (Lula, corrupto)*
"O Presidente viaja para Dinamarca e acompanha divulgação da sede das olimpíadas de 2016"	*viajar (Presidente,Dinamarca)* *acompanhar (Presidente, divulgação sede olimpíadas 2016)*

Para que seja possível construir uma *mr* a partir de uma oração é utilizado o PALAVRAS, um analisador sintático para língua portuguesa. A próxima subseção descreve como esse analisador funciona e, na seção seguinte o algoritmo que é capaz de criar uma *mr* a partir de uma oração.

4.4.2.2 PALAVRAS

O PALAVRAS (Bick, 2000) é um dos melhores analisadores sintáticos automáticos para o português do Brasil (MAZIERO, 2007). Dado um texto de entrada ele realiza a etiquetação sintática, léxica (palavras na forma canônica), e inclusive semântica para cada uma das palavras. A Figura 4.13 mostra um arquivo no formato XML (eXtensible Markup Language) exemplificando o retorno de uma análise feita pelo PALAVRAS.

```
<corpus id="sampleTIGER">
    <body>
<s id="s1" ref="1" source="Running text" forest="1" text="Lula ataca denuncismo e defende tratamento diferente para Sarney.">
    <graph root="s1_500">
        <terminals>
            <t id="s1_1" word="Lula" lemma="Lula" pos="prop" morph="M S" sem="--" extra="hum"/>
            <t id="s1_2" word="ataca" lemma="atacar" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="predco fmc mv"/>
            <t id="s1_3" word="denuncismo" lemma="denuncismo" pos="n" morph="M S" sem="activity" extra="--"/>
            <t id="s1_4" word="e" lemma="e" pos="conj-c" morph="--" sem="--" extra="co-fin co-fmc"/>
            <t id="s1_5" word="defende" lemma="defender" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="nosubj cjt-STA fmc mv"/>
            <t id="s1_6" word="tratamento" lemma="tratamento" pos="n" morph="M S" sem="activity" extra="--"/>
            <t id="s1_7" word="diferente" lemma="diferente" pos="adj" morph="M S" sem="--" extra="np-close"/>
            <t id="s1_8" word="para" lemma="para" pos="prp" morph="--" sem="--" extra="np-long"/>
            <t id="s1_9" word="Sarney" lemma="Sarney" pos="prop" morph="M/F S" sem="--" extra="hum"/>
            <t id="s1_10" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s1_500" cat="s">
                <edge label="STA" idref="s1_501"/>
            </nt>
            <nt id="s1_501" cat="fcl">
                <edge label="S" idref="s1_1"/>
                <edge label="X" idref="s1_502"/>
            </nt>
            <nt id="s1_502" cat="par">
                <edge label="CJT" idref="s1_503"/>
                <edge label="CO" idref="s1_4"/>
                <edge label="CJT" idref="s1_504"/>
            </nt>
            <nt id="s1_503" cat="x">
                <edge label="P" idref="s1_2"/>
                <edge label="Od" idref="s1_3"/>
            </nt>
            <nt id="s1_504" cat="x">
                <edge label="P" idref="s1_5"/>
                <edge label="Od" idref="s1_505"/>
            </nt>
            <nt id="s1_505" cat="np">
                <edge label="H" idref="s1_6"/>
                <edge label="DN" idref="s1_7"/>
                <edge label="DN" idref="s1_506"/>
            </nt>
            <nt id="s1_506" cat="pp">
                <edge label="H" idref="s1_8"/>
                <edge label="DP" idref="s1_9"/>
            </nt>
        </nonterminals>
    </graph>
</s>
```

Figura 4. 13. Exemplo do retorno de uma análise feita pelo PALAVRAS.

Observa-se que o retorno é no formato de uma árvore, onde há os nós terminais (t), marcados com s<numero da oração>_<unidade>, e os não terminais (nt), marcados com s<número da oração>_<centena> (Figura 4.13).

Os nós não terminais especificam a análise sintática da oração, por exemplo, o nó “s1_505” marcado como “od” (objeto direto) é composto pelos nós “s1_6” (tratamento), “s1_7” (diferente) e “s1_506”, que por sua vez é composto pelos nós “s1_8” (para) e “s1_9” (Sarney), isto é, “tratamento diferente para Sarney” (Figura 4.14). Já os nós terminais mostram, entre outras coisas, a especificação léxica e semântica das palavras.

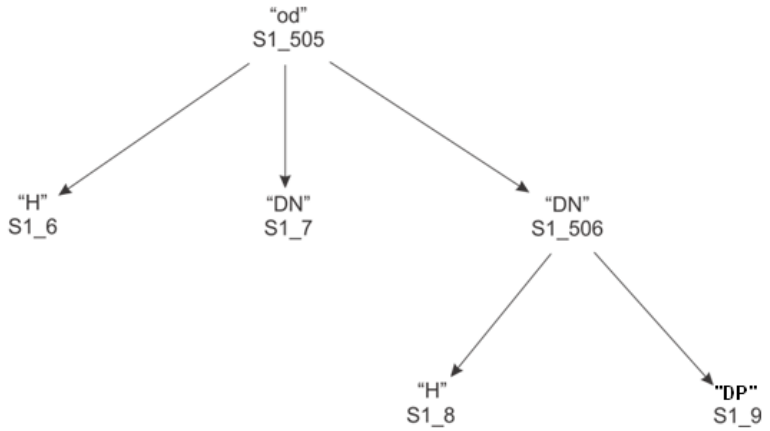


Figura 4. 14. Exemplo gráfico de uma análise sintática feita pelo PALAVRAS.

O PALAVRAS utiliza várias etiquetas para suas marcações, entretanto, esse trabalho considera apenas algumas delas que marcam os nós não terminais. Na Tabela 4.8 é apresentada a listagem (SYDDANSK UNIVERSITET, 2009).

Tabela 4. 8. Listagem das etiquetas usadas pelo PALAVRAS consideradas por este trabalho.

Etiqueta	**Significado**
adv	Advérbio
Ao	Complemento adverbial
art	Artigo
As	Complemento adverbial
CJT	Conjunto
Co	Predicativo do objeto
conj-c	Conjunção
conj-s	Conjunção
Cs	Predicativo do sujeito
fA	Adjunto adverbial
fCvo	Constituinte vocativo
intj	Interjeição
num	Numeral
od	Objeto direto (acusativo)
Odat	Objeto indireto pronominal
Oi	Objeto indireto pronominal
Op	Objeto preposicional
Opiv	Objeto preposicional
P	Predicador
pron-det	Pronome determinativo
pron-indp	Pronome independente
pron-pers	Pronome pessoal
prp	Preposição
pu	Pontuação
S	Sujeito
STA	Enunciado
UTT	Enunciado
X	Enunciado

Além do retorno em formato XML, o PALAVRAS disponibiliza outros, que não serão especificados por este trabalho, pois aqui apenas usa-se o PALAVRAS para a

análise sintática de uma oração para identificar qual é o *sujeito*, *verbo* e *complemento*, nada mais.

Uma explicação detalhada de cada uma das etiquetas usada por ele, além dos arquivos de retorno que ele disponibiliza, não estaria dentro do escopo deste trabalho. Por esse motivo, apenas as explicações dadas anteriormente são necessárias como base para o entendimento do método que aqui é proposto.

4.4.2.3 Como as metas-relação são construídas?

Primeiro, o algoritmo recebe uma oração de acordo com as especificações da seção 4.4.2.1, isto é, com sujeito, verbo e complemento. Para exemplificar será usada uma oração extraída de uma manchete de um jornal on-line: *"Lula ataca denuncismo e defende tratamento diferente para Sarney"*[18].

A oração é submetida ao analisador sintático PALAVRAS, que indica qual a categoria de cada um de seus componentes por meio de um arquivo ".xml", conforme discutido na seção anterior. Esse arquivo, por sua vez, é encaminhado a um algoritmo que divide a oração, caso seja necessário, e recupera o sujeito, verbo e o complemento da oração. No caso da oração usada como exemplo, a identificação de cada componente é a seguinte:

- Sujeito: *Lula;*
- Verbos: *ataca (1) – defende (2);*
- Complementos: *denuncismo (1) – tratamento diferente Sarney (2);*

O algoritmo é flexível o bastante para tratar as variações das orações, ou seja, dependendo há a possibilidade de recuperar mais de um verbo associado ao sujeito e mais de um complemento associado a um verbo. A Figura 4.15 ilustra um exemplo em que o sujeito (*Lula*) tem dois verbos associados, e para cada um dos verbos um complemento (*ataca – denuncismo; defende – tratamento diferente Sarney).*

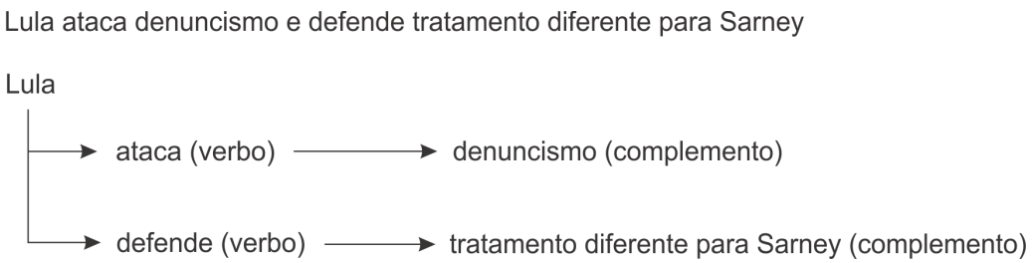


Figura 4. 15. Exemplo da atuação do algoritmo sobre uma análise feita pelo PALAVRAS.

[18] Manchete de uma notícia publicada pela Gazeta do Povo On Line no dia 17 de junho de 2009.

Para atender a essa flexibilidade foi definido um conjunto de regras que prevê como boa parte das orações construídas com sujeito, verbo e complemento são escritas. Essas regras foram definidas a partir de um estudo considerando mais de uma centena de orações.

Abaixo são listadas cada uma das regras com exemplos textuais e, como cada uma delas são interpretadas para gerar metas-relações:

- Regra 1 – para orações simples com sujeito, verbo e objeto:

  Estrutura da oração: <Sujeito> < verbo 1> < objeto 1>

  Estrutura da meta-relação: <verbo 1> (<sujeito>, <objeto 1>)

  Exemplo: "Lula defende Sarney" – defender (Lula, Sarney).

- Regra 2 – para orações com sujeito, verbo e objeto com complemento:

  Estrutura da oração: <Sujeito> <verbo 1> <objeto 1> <complemento 1>

  Estrutura da meta-relação: <verbo 1> (<sujeito>, <objeto 1> + <complemento 1>)

  Exemplo:

  "Lula defende tratamento diferenciado a Sarney"

  defender (Lula, tratamento diferenciado Sarney).

- Regra 3 – para orações com sujeito e uma conjunção separando dois verbos e dois objetos

  Estrutura da oração: <Sujeito> <verbo 1> <objeto 1> CONJ <verbo 2> <objeto 2>

  Estrutura da meta-relação:

  <verbo 1> (<sujeito>, <objeto 1>);

  <verbo 2> (<sujeito>, <objeto 2>)

  Exemplo:

  "Lula defende Sarney e ataca PSDB"

  defender (Lula, Sarney).

  atacar (Lula, PSDB).

- Regra 4 – para orações com sujeito, verbo e objeto mais complemento com uma conjunção separando um verbo e objeto.

  Estrutura da oração: <Sujeito> <verbo 1> <objeto 1> <complemento 1> CONJ <verbo 2> <objeto 2>

  Estrutura da meta-relação:

  <verbo 1> (<sujeito>, <objeto 1> + <complemento 1>)

<verbo 2> (<sujeito>, <objeto 2>)

Exemplo:

"Lula critica seqüência de denúncias sobre o Senado e defende Sarney"

criticar (Lula, sequencia denuncias Senado).

defender (Lula, Sarney).

- Regra 5 – para orações com sujeito, verbo e objeto mais complemento com uma conjunção separando um verbo, objeto mais complemento.

Estrutura da oração: <Sujeito> <verbo 1> <objeto 1> <complemento 1> CONJ <verbo 2> <objeto 2> <complemento 2>

Estrutura da meta-relação:

<verbo 1> (<sujeito>, <objeto 1> + <complemento 1>)

<verbo 2> (<sujeito>, <objeto 2> + <complemento 1>)

Exemplo:

"Lula defende tratamento diferenciado a Sarney e ataca PSDB de Serra."

defende (Lula, tratamento diferenciado Sarney).

atacar (Lula, PSDB Serra).

Para cada um dos componentes definidos nas estruturas das orações anteriormente, isto é, sujeito, verbo, objeto e complemento o algoritmo associa as respectivas etiquetas que o PALAVRAS utiliza para categorizá-los na análise sintática. Por exemplo, o sujeito é associado a etiqueta "S", o verbo a "P", etc. (Tabela 4.9).

Tabela 4. 9. Associação dos componentes de uma mr com as etiquetas usadas pelo PALAVRAS.

Componente	**Etiqueta relacionada**
<Sujeito>	S
<verbo>	P
<objeto>	Od, Oi
<complemento>	fA, Cs, S, fCvou, Oi, Odat, Opiv, Op, Opiv, As, Ao, Co, S

Essa associação se faz necessário para que o algoritmo consiga identificar cada um dos componentes da oração e como ela está estruturada, para que na construção da *mr* cada um deles seja acomodado em seu devido lugar, como apresentado na seguinte estrutura: <verbo> (<sujeito>, <objeto> + <complemento>).

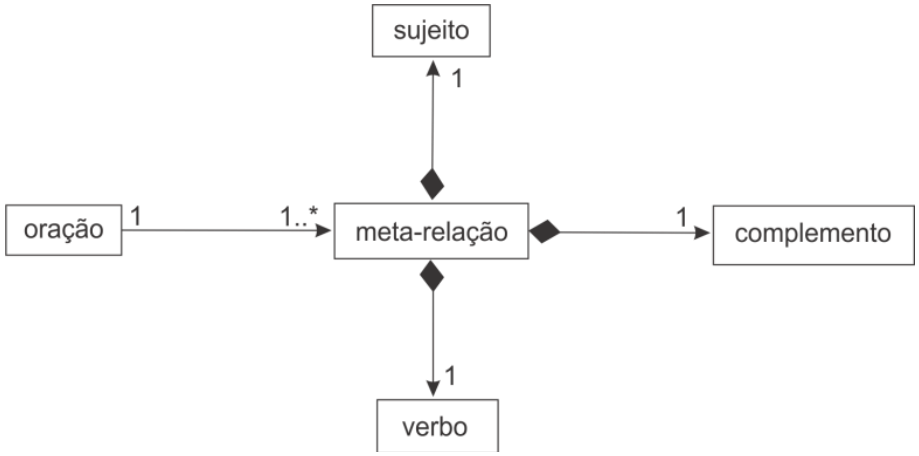


Figura 4. 16. Diagrama de classes representando uma *mr*.

Na Figura 4.16 é mostrado o modelo de classe que especifica as *mr*. Observa-se que uma oração pode gerar uma ou mais *mr*, como especificado nas regras 1, 2 e 3. Já na Figura 4.17 é apresentado o modelo gráfico, no qual representa a *mr* gerada da oração usada como exemplo. Observa-se que elas são construídas com a forma canônica das palavras.

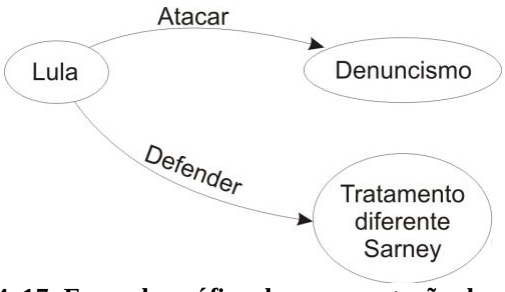


Figura 4. 17. Exemplo gráfico da representação de uma *mr*.

Finalmente, depois de um processamento em uma oração, que pode ser provida por um usuário ou retirada de qualquer lugar na web, conseguiu-se chegar a uma representação computacional (*mr*) capaz de armazenar a semântica de uma oração que possivelmente representa certo assunto.

Na próxima seção é apresentado como a base de conhecimento cultural do OMCS-Br é usada para enriquecer culturalmente o assunto representado por *concepts*, bem como a utilização da base de sinônimos; ambos são o último processo para formalizar *concepts* como uma representação de conhecimento.

4.4.2.4 Representação de conhecimento

Neste trabalho considera-se representação de conhecimento um conjunto de *mr*, oriundas de um conjunto de orações escrita em linguagem natural sobre um determinado assunto, que unidas são capazes de representar conhecimento relacionado a um assunto em questão.

A Figura 4.18 mostra uma representação de conhecimento sobre o assunto política brasileira. As orações usadas como sementes para essa representação são as seguintes: "*Lula ataca denuncismo e defende tratamento diferente para Sarney*" e "*Lula pede apuração correta e tratamento diferenciado a Sarney*".

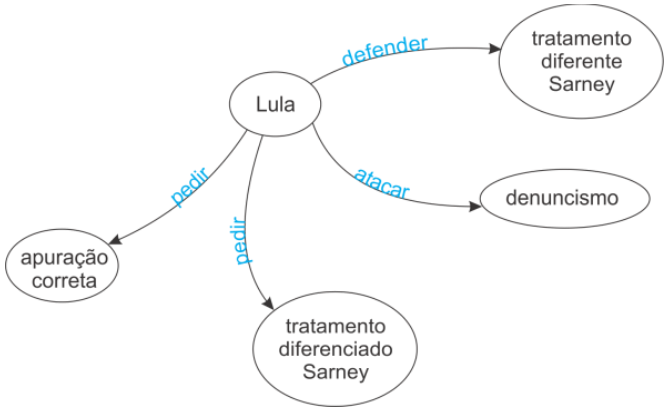


Figura 4. 18. Representação de conhecimento composta por um conjunto de mr geradas a partir de um conjunto de orações.

Tal representação de conhecimento é enriquecida com conhecimento cultural, originário da base do projeto OMCS-Br, a fim de adicionar maior representatividade ao assunto que ela mantém. Para que se consiga isso, são usadas as próprias *mr* que compõe a representação de conhecimento, ou seja, para cada uma são extraídos *s* e *c*, e usados como conceitos para uma busca de conhecimento cultural na base do OMCS-Br.

Para cada conhecimento com maior freqüência (veja na seção 3.4) conseguido, ou seja, dois conceitos relacionados por uma relação de Minsky, é recuperado a oração em linguagem natural que o originou.

Isso é possível porque toda relação semântica da base do OMCS-Br está relacionada a sentença em linguagem natural de onde ela foi extraída. A partir dessa oração é gerada uma nova *mr*, como discutido nas seções anteriores, na qual é adicionada a representação de conhecimento.

Para exemplificar esse processo considere *s = "Lula"*, que foi extraído de uma das *mr* da representação de conhecimento. Ao ser submetido à base do OMCS-Br ele recupera o seguinte conhecimento: *IsA (Lula, presidente, 18)*, que é procedente da sentença *"Lula é um presidente"*. Depois de processada, tal oração é transformada na *mr ser (Lula, presidente)*, que é adicionada a representação de conhecimento, como é mostrado na Figura 4.19.

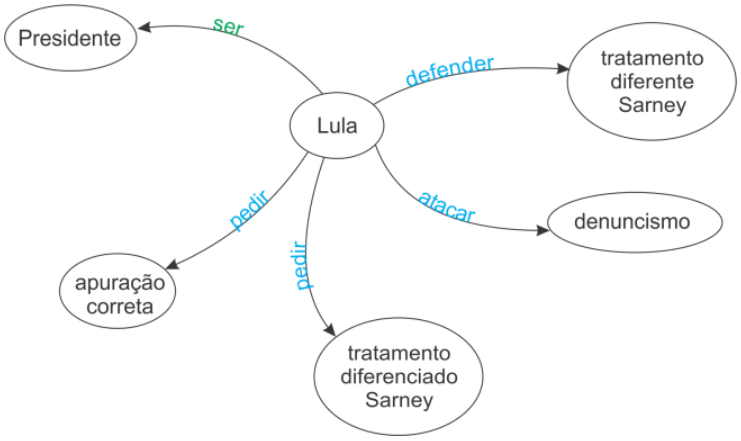


Figura 4. 19. Representação de conhecimento enriquecida com conhecimento cultural.

A representação de conhecimento, além de enriquecida com conhecimento cultural, pode ser expandida com o uso de sinônimos. Isso também é feito para melhorar a representatividade do assunto representado por ela.

Para isso, são usadas novamente as próprias *mr* que compõe a representação de conhecimento, como no processo de enriquecimento. Para cada uma é extraído *v* e submetido a um banco de sinônimos, provido pelo editor de texto Open Office. Caso é recuperado algum sinônimo relacionado a *v*, a *mr* em questão é duplicada e o verbo da segunda é substituído pelo verbo recuperado do banco de sinônimos.

Para exemplificar esse processo considere *v = "defender"*, que foi extraído da *mr defender (Lula, Sarney)*. Esse verbo conseguiu recuperar o sinônimo "*proteger*", então, cria-se uma meta-relação, *proteger (Lula, Sarney)*, e é adicionada a representação de conhecimento (Figura 4.20).

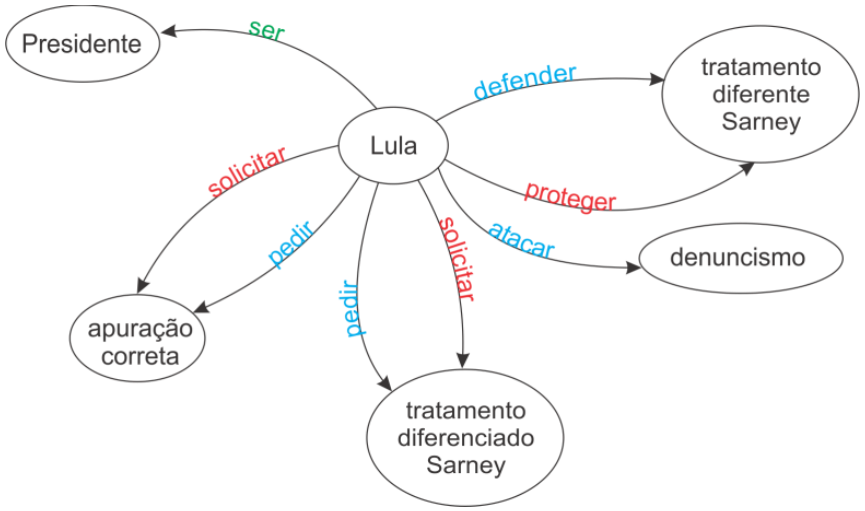


Figura 4. 20. Representação de conhecimento após o uso de banco de sinônimos.

Na Figura 4.20 é mostrado um exemplo da representação de conhecimento em sua completude. As metas-relações com os verbos em azul são oriundas das sentenças usadas

como sementes; as em verde são da base de conhecimento cultural do OMCS-Br e; as em vermelho são da base de sinônimos.

A próxima seção apresenta a evolução do método, na qual se usa todos os conceitos apresentados anteriormente.

4.4.2.5 Evolução do método.

A evolução do método teve como objetivo construir um conjunto de *mr* capaz de representar conhecimento contextualizado sobre determinado assunto e, posteriormente, identificar usuários de SNS que se expressam de acordo com tal conhecimento, com isso, consequentemente, são identificados usuários que estão falando sobre o mesmo assunto.

Inicialmente o algoritmo necessita de um conjunto de orações que representem determinado assunto. Para exemplificar serão usadas as seguintes orações: *"Lula defende Sarney" e "Lula ataca denuncismo e defende tratamento diferente a Sarney"*.

Cada uma das orações são submetidas ao PALAVRAS, que retorna um arquivo ".XML", como discutido na seção 4.4.2.2, com a análise sintática. O arquivo é submetido ao módulo de geração de *mr* (veja seção 4.4.2.3), que por sua vez retorna as *mr* conseguidas para cada uma das orações (Figura 4.21).

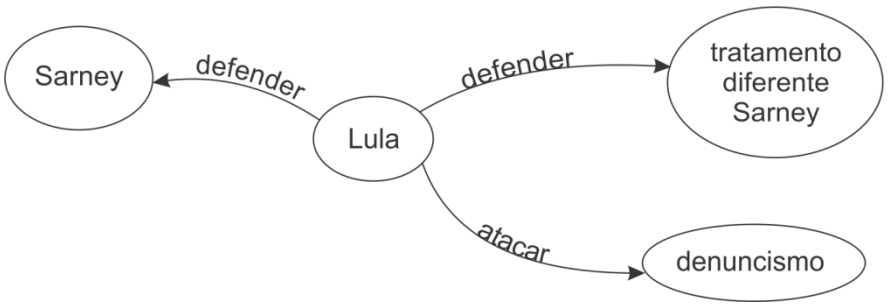


Figura 4. 21. . Relações semânticas geradas a partir de um conjunto de orações que expressam um assunto.

Esse processo consegue representar o conhecimento apenas das orações, por isso é necessário enriquecê-lo com conhecimento cultural e expandi-lo com o uso de um banco de sinônimos (a seção 4.4.2.4 mostra como é feito esse processo).

Finalmente, tem-se um conjunto de *mr* que representa conhecimento sobre certo assunto. A Figura 4.22 mostra o final de todo o processo, em destaque estão as *mr* geradas com o auxílio da base de conhecimento cultural e o banco de sinônimos.

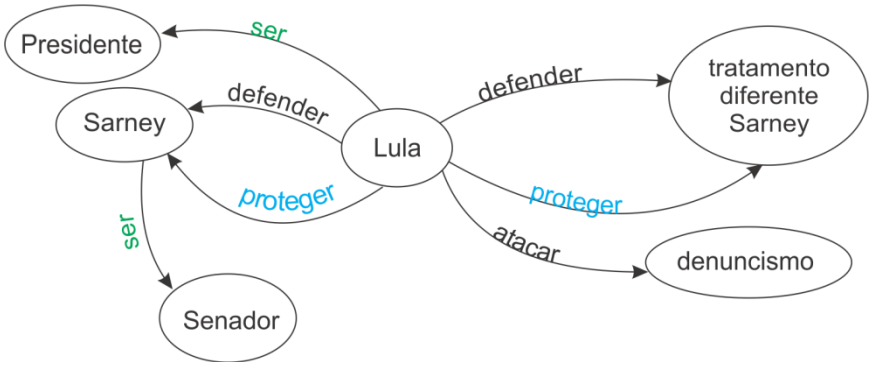


Figura 4. 22. Parte de representação de conhecimento sobre um assunto.

A busca por usuários em SNS é similar a versão anterior do método. São usados pares de palavras para as buscas nos Fóruns das Comunidades do Orkut. Tais pares de palavras são oriundos de cada *mr*. Por exemplo, para a *mr defender (Lula, Sarney)* são usados apenas *s* (*Lula*) e *c* (*Sarney*) com exceção de *v* (*defender*).

Para essa busca é usada todas as *mr* da representação de conhecimento com seus respectivos pares de palavras. A Figura 4.23 mostra o resultado de uma busca.

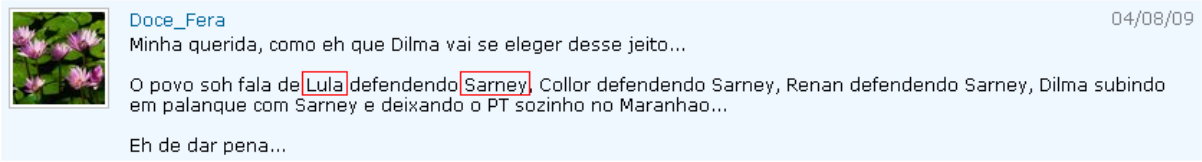

Doce_Fera 04/08/09
Minha querida, como eh que Dilma vai se eleger desse jeito...
O povo soh fala de Lula defendendo Sarney, Collor defendendo Sarney, Renan defendendo Sarney, Dilma subindo em palanque com Sarney e deixando o PT sozinho no Maranhao...
Eh de dar pena...

Figura 4. 23. Exemplo de uma busca usando uma mr pertencente a representação de conhecimento.

Cada postagem identificada é armazenada em um banco de dados associada a respectiva *mr* usada para a busca, para uma posterior verificação. Depois de todas as postagens armazenadas, cada uma delas passa por um novo processo que visa identificar o par de palavras que a identificou em um mesmo trecho de texto. Por exemplo, a postagem da Figura 4.23 foi identificada pelo par de palavras *"Lula"* e "*Sarney"*, percebe-se que entre elas não há nenhum tipo de pontuação.

Depois dessa verificação o trecho de texto que está entre as duas palavras e as mesmas são obtidas da postagem (Figura 4.24).

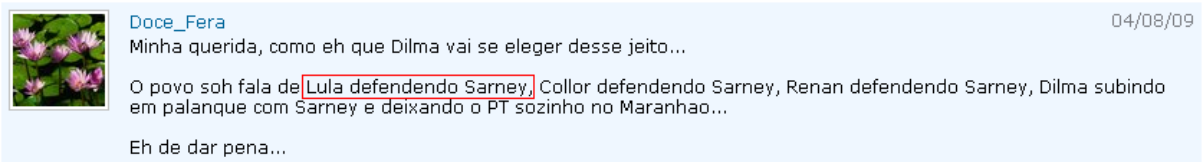

Doce_Fera 04/08/09
Minha querida, como eh que Dilma vai se eleger desse jeito...
O povo soh fala de Lula defendendo Sarney, Collor defendendo Sarney, Renan defendendo Sarney, Dilma subindo em palanque com Sarney e deixando o PT sozinho no Maranhao...
Eh de dar pena...

Figura 4. 24. Exemplo de recuperação do trecho de texto de uma postagem que possivelmente, depois de transformado em mr, pode ser igual a mr em questão usada na busca.

Finalmente, o trecho de texto removido passa por outro processo no qual é construído uma *mr*, que é comparada com a *mr* usada no processo de busca da postagem. Por

exemplo, o trecho de texto "*Lula defendendo Sarney*" retirado da postagem da Figura 4.24 gera a *mr defender (Lula, Sarney)* igual a *mr* usada na busca.

Caso ocorra a igualdade, como no exemplo anterior, o link do usuário responsável pela postagem é recuperado e adicionado ao conjunto β, o mesmo da versão anterior do método.

Com isso possivelmente pode se dizer que os usuários selecionados se expressam da mesma forma que parte da representação de conhecimento, além disso, há indícios de que eles estão falando sobre o mesmo assunto e possivelmente podem possuir o mesmo consenso.

A representação de conhecimento com um conjunto de *mr* é justificada pelo seguinte fato. As duas orações usadas no exemplo "*Lula defende Sarney*" e "*Lula defendendo Sarney*" são semanticamente iguais, mas por uma comparação textual elas são consideradas distintas. Com o uso de *mr* as duas orações passam a ser consideradas iguais, pois elas são representadas da seguinte forma: *defender (Lula, Sarney).*

Por esse motivo que os trechos de textos encontrados nas postagens são transformados em *mr*, pois, dessa forma é possível considerar todo o tipo de variação temporal entre as orações. Orações como *"Lula defende Sarney", "Lula defendia Sarney", "Lula defenderá Sarney",* etc. são consideradas iguais, uma vez que para transformar em *mr* todas palavras ficam em sua forma canônica.

Isso mostra que o método dessa vez tenta considerar melhor a semântica envolvida entre as palavras, e não somente comparação textual como nas versões anteriores e, também como nos trabalhos de Chen (2009) e Granada *et. al* (2006).

4.4.3 Teste

Foi conduzido um experimento para verificar se a abordagem adotada na atual versão do método supera a versão anterior. Os passos adotados para isso são:

(1) Escolha das orações que representam um assunto;
(2) Representação de conhecimento em relação ao assunto usando mr, ou seja, construção de concepts e;
(3) Desenvolvimento de um aplicativo que busca postagens no Orkut e as compara com as mr de concepts;

As orações usadas como sementes para a construção de *concepts* foram extraídas das manchetes de três portais de notícias on-line do Brasil, ou seja, as mesmas usadas no experimento da versão anterior. A saber: *"Lula critica sequência de denúncias sobre o Senado e defende Sarney", "Lula defende Sarney e diz que denúncias não têm fim"* e *"Lula pede apuração correta e tratamento diferenciado a Sarney"*.

Após o processamento das orações têm-se um conjunto de *mr* que, nesse caso, representa o assunto política. A Tabela 4.10 mostra as *mr* correspondentes de cada oração*;* A Tabela 4.11 mostra a nova *mr* que cada uma conseguiu com o auxílio da base de conhecimento cultural e; finalmente a Tabela 4.12 mostra o que cada *mr* conseguiu com o auxílio do banco de sinônimos.

Tabela 4. 10. mr oriundas das orações usadas como sementes.

Oração	***Mr***
Lula critica sequência de denúncias sobre o Senado e defende Sarney	*criticar (Lula, sequência denúncias senado)* *defender (Lula, Sarney)*
Lula defende Sarney e diz que denúncias não têm fim	*defender (Lula, Sarney)* *dizer (Lula, denúncias não têm fim)*
Lula pede apuração correta e tratamento diferenciado a Sarney	*pedir (Lula, apuração correta)* *pedir (Lula, tratamento diferenciado Sarney)*

Tabela 4. 11. *mr* geradas a partir do auxílio da base de conhecimento cultural do projeto OMCS-Br.

Mr	***Mr* extraída da base do OMCS-Br**
criticar (Lula, sequência denúncias senado)	*ser (Lula, Presidente)*
defender (Lula, Sarney)	*ser (Lula, Presidente); ser (Sarney, Senador)*
dizer (Lula, denúncias não têm fim)	*ser (Lula, Presidente)*
pedir (Lula, apuração correta)	*ser (Lula, Presidente)*
pedir (Lula, tratamento diferenciado Sarney)	*ser (Lula, Presidente)*

Tabela 4. 12. mr geradas a partir do uso da base de sinônimos.

Mr	***Mr* extraída da base de sinônimos**
criticar (Lula, sequência denúncias senado)	Não encontrou sinônimo
defender (Lula, Sarney)	*proteger (Lula, Sarney)*
dizer (Lula, denúncias não têm fim)	*falar (Lula, denúncias não têm fim)*
pedir (Lula, apuração correta)	*solicitar (Lula, apuração correta)*
pedir (Lula, tratamento diferenciado Sarney)	*solicitar (Lula, tratamento diferenciado Sarney)*

Observa-se na Tabela 4.11 que foi encontrado apenas conhecimento cultural sobre as palavras "*Lula*" e "*Sarney*". Esse fato é atribuído a forma como foi conduzida a busca por conhecimento na base do OMCS-Br. Por exemplo, quando se utilizou *s* da *mr* a busca considerou apenas uma palavra - "*Lula*"; quando se utilizou *c* da *mr* a busca, com exceção da palavra "*Sarney*" considerou sempre mais de uma palavra – "*sequência denúncias*

senado", etc. Isso prejudica muito os resultados, pois é difícil encontra conhecimento cultural com uma busca usando agrupamento de substantivos, como nesse caso.

Finalmente, tem-se o conjunto de *mr* formado pelas *mr* extraídas das orações, da base de conhecimento cultural e do banco de sinônimos, isto é, uma representação de conhecimento. Para usar as *mr* na busca por postagens, foi usada basicamente a mesma estratégia da versão anterior do método, ou seja, foi desenvolvido um aplicativo, baseado em um analisador de HTML, que vasculha a estrutura das páginas dos fóruns de discussão das Comunidades do Orkut. Cada postagem potencial encontrada, ou seja, que possui *s* e *c* da *mr* usada na busca, são armazenadas em um banco de dados associada a *mr* em questão. Os resultados do experimento serão descritos na próxima seção.

4.4.3.1 Resultados

O saldo total das postagens recuperadas dos fóruns das comunidades do Orkut foi de 5.779. 1.76% (102) tinham pelo menos uma *mr* igual a uma das *mr* que representam o conhecimento usado como parâmetro na busca.

Na Tabela 4.13 é mostrado o quanto cada *mr* conseguiu recuperar de postagens com trechos de textos, depois de processados, iguais semanticamente as *mr* da representação de conhecimento, geradas a partir das orações usadas como semente.

Tabela 4. 13. Listagem do desempenho de cada mr geradas a partir do uso da oração, base cultural do OMCS-Br e da base de sinônimos.

mr	**Quantidade de postagens**	**(%)**
criticar (Lula, sequência denúncias senado)	4	3.92
defender (Lula, Sarney)	81	79.41
dizer (Lula, denúncias não têm fim)	0	0
pedir (Lula, apuração correta)	0	0
pedir (Lula, tratamento diferenciado Sarney)	2	1.96
proteger (Lula, Sarney)	4	3.92
falar (Lula, denúncias não têm fim)	0	0
solicitar (Lula, apuração correta)	0	0
solicitar (Lula, tratamento diferenciado Sarney)	0	0
ser (Lula, Presidente)	6	5.88
ser (Sarney, Senador)	5	4.91

Na Figura 4.25 é mostrado o quanto as *mr* oriundas das orações, da base de conhecimento cultural e do banco de sinônimos conseguiram identificar individualmente trechos de textos iguais semanticamente a elas.

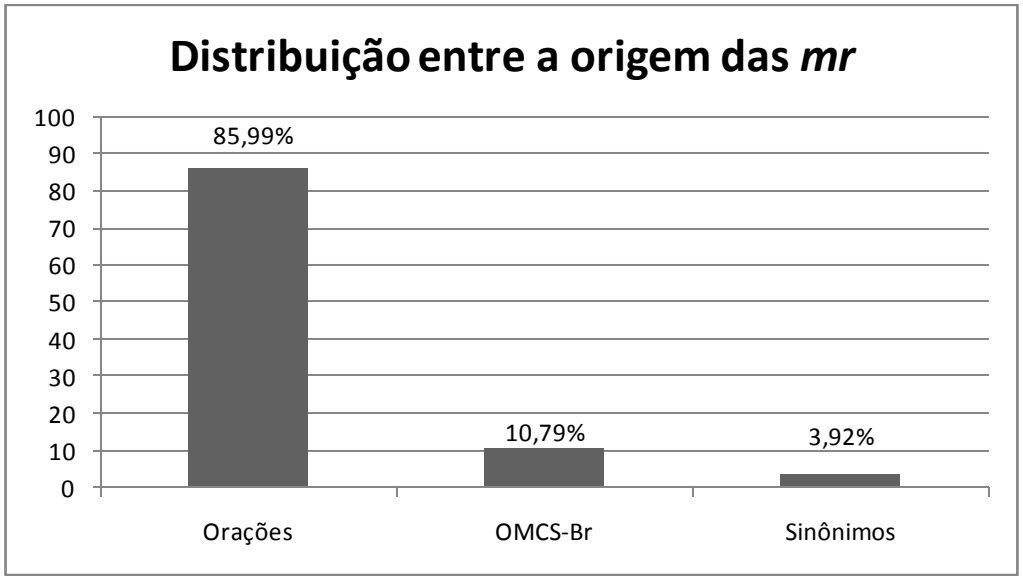


Figura 4. 25. Relação entre o desempenho das mr geradas a partir das orações com o apoio da base do OMCS e do apoio da base de sinônimos.

Finalmente o conjunto β conseguiu um total de 26 usuários, sendo 11 recuperados pela *mr defender (Lula, Sarney).*

4.4.4 Apontar falhas

Na versão atual do método houve uma melhora na quantidade de usuários selecionados em relação à versão anterior. Isso se dá pela forma de busca usada para recuperar as postagens, pois o assunto dessa vez era o mesmo.

Na versão anterior era mais difícil encontrar o conjunto de palavras do vetor *concepts* usando como referência apenas duas palavras. Nessa versão as chances aumentam devido à busca ser apenas pela semântica envolvida entre as duas palavras, ou seja, não depende de outras palavras como no anterior.

Uma falha identificada nessa versão é referente também ao falso-positivo, pois as *mr* geradas com o auxílio da base do OMCS-Br têm muito pouco significado em relação ao assunto. O trecho de texto *"Lula é um presidente que gosta de jogar futebol...."* recuperado de uma postagem pela *mr ser (Lula, presidente)* é um exemplo de falso-positivo identificado durante o experimento.

Percebe-se que a postagem não tem relação alguma com o assunto usado no experimento, mas seu responsável, ou melhor, o usuário, foi selecionado como uma pessoa que está falando sobre o assunto.

Em relação as *mrs* oriundas diretamente das orações e da base de sinônimos não houve problemas em relação a falso-positivo, pois, todos os trechos das postagens dos

usuários que foram selecionadas pelas *mr* como semanticamente similares aparentemente estavam dentro do assunto.

O desempenho das *mr* oriundas da base de sinônimos não apresentou resultado satisfatório, pois houve apenas 3.92% de sucesso. Talvez os sinônimos usados para a geração das *mr* não são frequentemente usados pelas pessoas em seus cotidianos.

Todos esses dados e comentários levam a conclusão que se conseguiu um bom avanço em relação às versões anteriores do método, pois, se conseguiu identificar usuários em SNS que realmente estão falando sobre certo assunto, entretanto, ainda há a necessidade de se aproveitar melhor o potencial que a base do OMCS-Br tem para prover conhecimento cultural.

4.5 Iteração 4 – Identificando pessoas que falam sobre o mesmo assunto tornando as diferenças culturais irrelevantes.

Na iteração anterior percebeu-se uma melhora significativa do método, mas ainda são percebidos alguns problemas, como a forma como a base de conhecimento cultural do projeto OMCS-Br está sendo usada, além disso, a tentativa de utilizar a base de sinônimos para aumentar o alcance da busca não apresentou resultados satisfatórios.

Devido a isso, a tentativa de melhorar o método nessa iteração se dá em dispensar o uso da base de sinônimos e, usar a base do projeto OMCS-Br como uma fonte de sinônimos cultural.

Essa iteração é a última instanciada para o refinamento do método. Ela é de fato a proposta deste trabalho, ou seja, um método que identifica usuários de redes sociais que estão falando sobre o mesmo assunto.

4.5.1 Exposição do problema

A versão atual do método define um conjunto de relações semânticas (*mr*) que representam um assunto. A esse conjunto dá-se o nome de representação de conhecimento, que é usado como parâmetro para identificar usuários em SNSs que estão falando sobre o mesmo assunto.

Essa solução se mostrou eficiente no que diz respeito ao melhor relacionamento semântico entre as palavras usadas para representar um assunto. Porém, alguns pontos apresentaram problemas, como a forma de uso de conhecimento cultural provido pela base do OMCS-Br.

O uso de banco de sinônimos não apresentou bons resultados, porém, serviu para chamar atenção sobre o vocabulário que as pessoas usam em SNSs, pois, os sinônimos formais usados para expandir a representação de conhecimento, frequentemente não são usados pelas pessoas.

Dessa forma, o problema a ser resolvido por essa iteração está relacionado a um melhor uso da base de conhecimento cultural do OMCS-Br, além de dispensar o uso de base de sinônimos.

4.5.2. Resolução do problema

Com o experimento realizado na iteração anterior foi possível observar que os usuários de SNSs, ou pelo menos do Orkut, utilizam um modo mais informal e diversificado de se expressarem, como o uso de analogias, sinônimos oriundos da cultura popular, do regionalismo, etc.

Esse tipo de comportamento é a forma de expressão da cultura das pessoas, e é similar a expressão de conhecimento cultural dos colaboradores do projeto OMCS-Br quando fornecem seus conhecimentos ao projeto.

Por exemplo, em uma simples consulta na base de conhecimento a fim de conseguir sinônimos de um conceito, é possível ter grandes surpresas com o retorno de palavras que não fazem parte de nosso vocabulário, mas em outra região do país, ou melhor, em outra cultura, são palavras usadas normalmente. Por exemplo, a palavra “mandioca” pode ser desconhecida no nordeste do país, onde seu sinônimo é “macaxera”.

Essa diversidade é justificada pela grande heterogeneidade no perfil dos colaboradores do projeto, pois a faixa etária, a localização e até mesmo o grau de escolaridade pode influenciar na cultura das pessoas, principalmente no vocabulário.

Na Tabela 4.14, Tabela 4.15 e Tabela 4.16 são apresentados os dados referentes a faixa etária, distribuição geográfica e escolaridade dos colaboradores do projeto OMCS-Br, respectivamente, no qual justifica a diversidade cultural da base.

Tabela 4. 14. Distribuição da faixa etária dos colaboradores do projeto OMCS-Br.

Faixa etária	Quantidade	Porcentagem
12 anos	11	0,64
Entre 13 e 17	239	13,94
Entre 18 e 29	1128	65,81
Entre 30 e 45	272	15,87
Entre 46 e 65	64	3,73
Mais de 65	10	0,58

Tabela 4. 15. Distribuição dos colaboradores do projeto OMCS-Br por região.

Estado	Quantidade	Porcentagem
AC	4	0,24
AL	5	0,30
AM	11	0,67
BA	25	1,52
CE	30	1,83
DF	27	1,65
ES	39	2,38
GO	27	1,65
MA	3	0,18
MG	70	4,27
MS	16	0,98
MT	7	0,43
PA	9	0,55
PB	13	0,79
PE	28	1,71
PI	7	0,43
PR	85	5,18
RJ	74	4,51
RN	4	0,24
RO	1	0,06
RR	1	0,06
RS	103	6,28
SC	65	3,96
SE	5	0,30
SP	975	59,45
TO	6	0,37

Tabela 4. 16. Distribuição dos colaboradores do projeto OMCS-Br por nível de escolaridade.

Nível de escolaridade	Quantidade	Porcentual
Ensino Fundamental	130	7,10
Ensino Médio	842	49,23
Graduação	433	25,38
Pós-Graduação	104	5,92
Mestrado	156	8,88
Doutorado	59	3,49

Da mesma forma que o projeto OMCS-Br, o Orkut também possui usuários de todas regiões do país, assim, a mesma diversidade cultural, principalmente de vocabulário, que é percebida na base do OMCS-Br, também pode ser percebida nas interações sociais dos usuários no Orkut. Por exemplo, algumas pessoas no Orkut se referem à cidade de São Paulo, além de seu nome, como "terra da garoa"; a cidade do Rio de Janeiro, como "cidade maravilhosa"; etc.

Esses tipos de sinônimos, que neste trabalho são considerados culturais, fez com que houvesse a percepção de um melhor uso do potencial da base de conhecimento cultural, ou seja, como um banco de sinônimos cultural, pois, as pessoas transferem seus vocabulários para as interações sociais nos SNS, e na base de conhecimento cultural está uma grande amostra da representação do vocabulário dessas pessoas.

Com isso é possível encontrar pessoas em SNS que estão falando sobre o mesmo assunto, mas que se expressam de forma diferente. Por exemplo, as duas frases ditas por pessoas diferentes "O Rio de Janeiro continua lindo" e "Cidade maravilhosa permanece bela" expressam a mesma coisa, mas são escritas de forma diferente. Para identificar esse tipo de similaridade é que se pretende evoluir o método.

Para atingir esse objetivo a base de conhecimento cultural será utilizada também como fonte de sinônimos cultural, ou seja, fonte de vocabulário específico que só existe na cultura das pessoas, e que não é encontrado em dicionários de sinônimos algum. Isso possivelmente melhorará a qualidade das buscas, podendo até identificar grupos de pessoas que têm um vocabulário particular quando se fala sobre certo assunto.

Essa iteração se baseia na seguinte regra:

(4) Se duas pessoas falam as mesmas coisas, independente das diferenças culturais expressas pelo vocabulário, elas estão falando sobre o mesmo assunto.

Nas próximas subseções são apresentados alguns conceitos e abordagens adotadas por este trabalho. Na seção 4.5.2.1 e 4.5.2.2 são apresentados os conceitos sobre as relações de Minsky aptas a representar variações de vocabulário. Na seção 4.5.2.3 é apresentado como a base de conhecimento cultural é usada como uma base de sinônimo cultural; e finalmente na seção 4.5.2.4 é mostrada a evolução do método, que é tida como a proposta deste trabalho.

4.5.2.1 IsA

IsA é considerada uma relação fraca (LIU, 2004) e seu propósito é especializar algo hierarquicamente. Ela é representada da seguinte forma: *IsA(X, Y),* onde *X* é um conceito especializado a partir do conceito genérico *Y.* O que se quer dizer é que *X,* além de possuir todas as características de *Y,* possui no mínimo uma característica a mais, o que o faz derivado de *Y.*

Por exemplo, *IsA(Rio de Janeiro, cidade) (Rio de Janeiro é uma cidade),* todas as características que uma "*cidade"* possui "*Rio de Janeiro"* também possui, porém "*Rio de Janeiro"* possui características adicionais como *"possuir o pão de açúcar"*. A Figura 4.26 mostra a representação gráfica da relação.

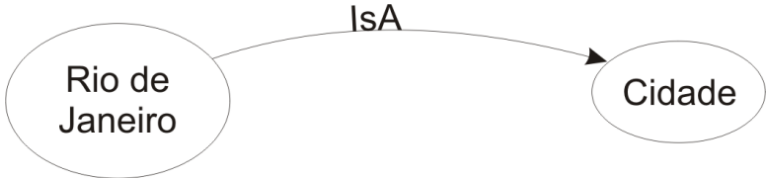


Figura 4. 26. Representação gráfica da relação IsA.

Levando em consideração as explicações anteriores, este trabalho faz a seguinte definição:

(1) Se *X* é uma especialização de *Y,* então as *caract (Y)* ⊂ *caract(X).*

Essa definição afirma que *(X → Y),* isto é, ***o conceito representado por Y pode ser usado para representar X***, mas não o inverso. Por exemplo, quando se tem o conhecimento *IsA(Rio de Janeiro, cidade),* pode-se referir a *"Rio de Janeiro"* usando a palavra *"cidade"* (*Rio de Janeiro → cidade*), mas não se referir a "*cidade*" qualquer usando a palavra "*Rio de Janeiro*".

4.5.2.2 DefinedAs

DefinedAs é um tipo de relação que faz uso de sinônimos para representar o significado de algo (LIU, 2004). Ela é representada da seguinte forma: *DefinedAs(X, Y)*, onde *X* é uma conceito que tem a mesma natureza que o conceito *Y*. O que se quer dizer é que *X* possui todas as características de *Y* e vice-versa. Por exemplo, *DefinedAs(Linda, Maravilhosa) (Linda é definido como Maravilhosa),* todas as características que *"Linda"* possui *"Maravilhosa"* também possui. A Figura 4.27 mostra a representação gráfica da relação.

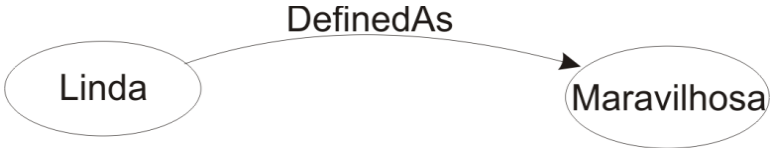


Figura 4. 27. Representação gráfica da relação DefinedAs.

Desta forma é possível fazer a seguinte definição:

(2) Se X é sinônimo de Y, então caract (X) ⇔ caract(Y).

Essa definição garante que as *caract (X) ⇔ caract(Y),* isso remete a *(X ↔ Y),* isto é, ***o conceito representado por X pode ser usado para representar Y e vice-versa***. Por exemplo, quando se tem o conhecimento *DefinedAs(Linda, Maravilhosa),* pode-se referir a "*Maravilhosa*" usando a palavra "*Linda*" e vice-versa.

4.5.2.3 Base do OMCS-Br como fonte de sinônimo cultural

O uso da base de conhecimento cultural como uma fonte de sinônimo é visto como uma forma de expansão semântica de conhecimento. Considera-se expansão de conhecimento o fato de poder derivar uma *mr* (representação de conhecimento) em um conjunto $\alpha = \{mr_1...mr_n\}$ de outras novas metas-relação, onde cada mr_n possui o mesmo significado e valor semântico que a *mr* base sem perder o assunto e contexto em questão.

Na Tabela 4.17 são apresentados alguns exemplos de expansão semântica de *mr* distintas.

Tabela 4. 17. Exemplo de expansão semântica de metas-relação.

***mr* base**	***mr*$_n$**
continuar (Rio de Janeiro, lindo)	*permanecer(Rio de Janeiro, lindo)* *continuar (Cidade Maravilhosa, lindo)*
destruir (Terremoto, Chile)	*devastar (Terremoto, Chile)* *destruir (Tremor, Chile)* *destruir (Terremoto, país Andino)*
defender(Lula, Sarney)	*proteger(Presidente, Sarney)* *defender(Presidente, Sarney)*

	proteger(Lula, Sarney)

A *mr = continuar (Rio de Janeiro, lindo)* na Tabela 4.17 foi extraída da oração "*O Rio de Janeiro continua lindo"*. Quando se faz processo inverso com a mr_1 *= permanecer (Rio de Janeiro, lindo)* derivada da *mr* base (Tabela 4.17) obtém-se *a* oração *"O Rio de Janeiro permanece lindo"*. Com isso observa-se que as duas metas-relação (*mr* e mr_1) são similares quanto ao significado e assunto, variando apenas culturalmente, demonstrando a principal finalidade da expansão de conhecimento.

Quando se realiza uma busca de conhecimento cultural na base do *OMCS-Br* é utilizado um conceito como referência, como exemplificado na seção 3.1.4, então, todo o conhecimento relacionado ao conceito em questão é recuperado da base.

Por exemplo, com a submissão do conceito "*Rio de Janeiro*" pode haver um retorno de conhecimento cultural semelhante a Figura 4.28 (a), além disso, é possível fazer uma busca associando um conceito a uma relação, como por exemplo, *DefinedAs(Rio de Janeiro, Y)*. Dessa forma, tudo que está relacionado com "*Rio de Janeiro*" por meio da relação *DefinedAs* é recuperado da base, como apresentado na Figura 4.28 (b).

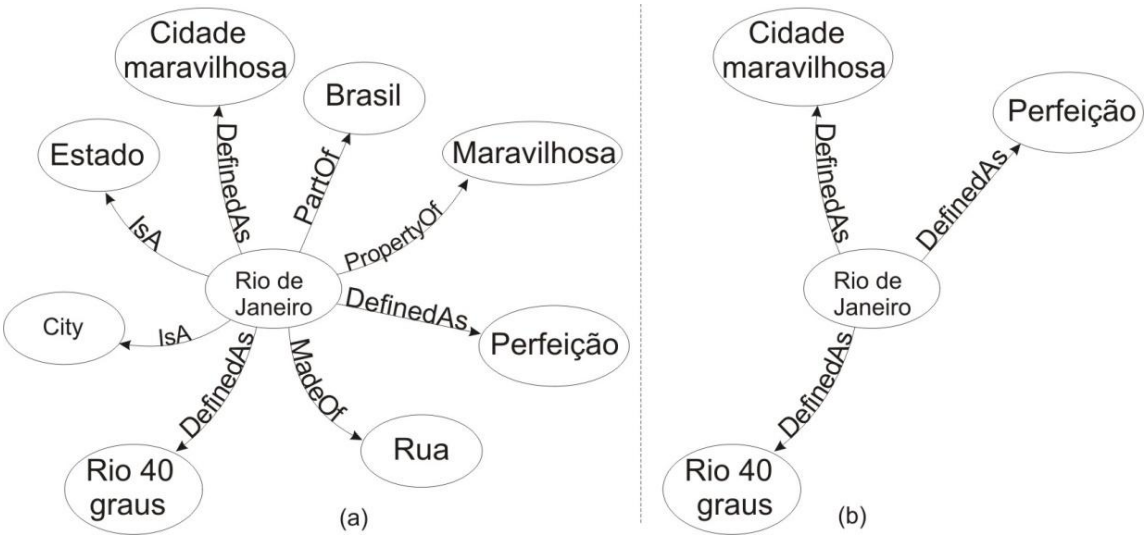


Figura 4. 28. Dois tipos possíveis de busca de conhecimento cultural na base do OMCS-Br.

Para a expansão de conhecimento é utilizada a segunda forma de busca, ou seja, um conceito associado com uma relação, além disso, são utilizadas apenas as relações *Isa* e *DefinedAs*, porque são as únicas relações do *OMCS-Br* que têm potencial em representar variações da língua, como mostrado nas definições (1) e (2) anteriormente.

Os parâmetros utilizados no processo de busca são os componentes de uma *mr (verbo, sujeito* e *objeto),* como apresentado na Figura 4.29. Por exemplo, *DefinedAs({verbo, sujeito ou objeto }, X).*

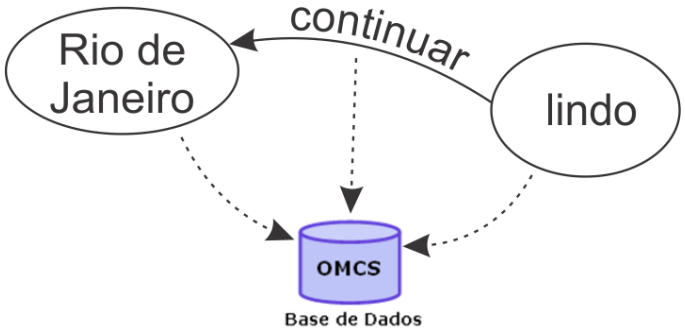


Figura 4. 29. Submissão dos componentes de mr como referência ao OMCS-Br, a fim de identificar sinônimos culturais.

Para uma busca coerente ao objetivo, que é identificar conhecimento cultural que possa ser usado como sinônimo cultural, há alguns critérios que devem ser considerados no uso dos componentes da *mr* como parâmetros:

- Quando se usa o *sujeito*: *IsA(sujeito, Y), DefinedAs(X, sujeito)* e *DefinedAs(sujeito, Y)*;
- Quando se usa o *complemento*: *IsA(complemento, Y), DefinedAs(X, complemento)* e *DefinedAs(complemento, Y);*
- Quando se usa o *verbo*: *DefinedAs(X, verbo) e DefinedAs(verbo, Y);*

A relação *IsA* é a mais restrita entre as duas relações, quando ela é usada a parte *X* é sempre fixa, pois como definido anteriormente, se $(X \rightarrow Y)$ pode-se usar apenas a palavra em *Y* para representar a palavra em *X*, mas não o inverso.

Por exemplo, com a busca *IsA(Rio de Janeiro, Y)* tem-se o *Y = "cidade"* como retorno, a palavra "*cidade"* pode ser usada para referir-se a palavra "*Rio de Janeiro"*, mas a palavra "*Rio de Janeiro"* não pode ser usado para referir-se a "*cidade",* pois quando aponta para "*Rio de Janeiro"* dizendo que é uma "*cidade"* não causa impacto quanto ao significado, mas quando aponta para uma "*cidade"* qualquer dizendo que é "*Rio de Janeiro"* há uma grande incoerência.

Pode-se observar também que quando se usa o *verbo,* a busca por conhecimento fica restrita apenas a relação *DefinedAs,* porque não existe representação hierárquica entre verbos. Para exemplificar o processo de busca considere *mr = continuar (Rio de Janeiro, lindo).* Na Tabela 4.18 é mostrado os parâmetros de busca e os resultados obtidos da base do *OMCS-Br* usando *IsA* e *DefinedAs*.

Tabela 4. 18. Resultado da busca na base de conhecimento cultural do OMCS-Br usando as relações IsA e DefinedAs.

Parâmetros de busca	Resultados em *X* ou *Y*
IsA (Rio de Janeiro, Y)	*Cidade, capital, cidade perigosa*
IsA (lindo, Y)	*Não houve resultado*
DefinedAs(X, Rio de Janeiro)	*Samba, cidade do Brasil, praia, beleza.*
DefinedAs(Rio de Janeiro, Y)	*Cidade maravilhosa, Rio, cidade maravilhosa.*
DefinedAs(X, lindo)	*Bonito.*
DefinedAs(lindo, Y)	*belo, maravilhoso*
DefinedAs(X, continuar)	*Permanecer, prosseguir*
DefinedAs(continuar, Y)	*Prosseguir, seguir, manter*

Os resultados obtidos na busca, de acordo com a Tabela 4.18, são usados como sinônimos culturais para expandir semanticamente a *mr = continuar (Rio de Janeiro, lindo).* Cada conceito obtido no resultado pode ser usado para substituir o respectivo componente (*verbo, sujeito* ou *objeto*) da *mr* usado como parâmetro na busca, resultando em uma nova *mr*. Por exemplo, o conceito "*Cidade maravilhosa"* obtido na busca por *DefinedAs(Rio de Janeiro, Y)* pode substituir o componente "*Rio de Janeiro" (sujeito)* na *mr = continuar (Rio de Janeiro, lindo),* derivando-a para mr_1 *= continuar (Cidade maravilhosa, lindo),* pois na busca *DefinedAs({sujeito, objeto ou verbo}, Y)* adota-se *(X ↔ Y),* isto é , nesse caso *Y* pode ser usado para representar *X*.

São geradas novas *mr* por permutação, ou seja, pode-se derivar outra *mr* a partir de mr_1, por exemplo, a mr_1 *= continuar (Cidade maravilhosa, lindo)* geraria *permanecer (Cidade maravilhosa, lindo),* nesse caso "*permanecer"* substitui "*continuar",* pois a busca *DefinedAs(X, continuar)* obteve-se o conceito "*permanecer".* Deixa-se claro que a permutação apenas é possível entre os mesmos componentes da *mr,* isto é, *sujeito* com *sujeito, verbo* com *verbo* e *complemento* com *complemento*.

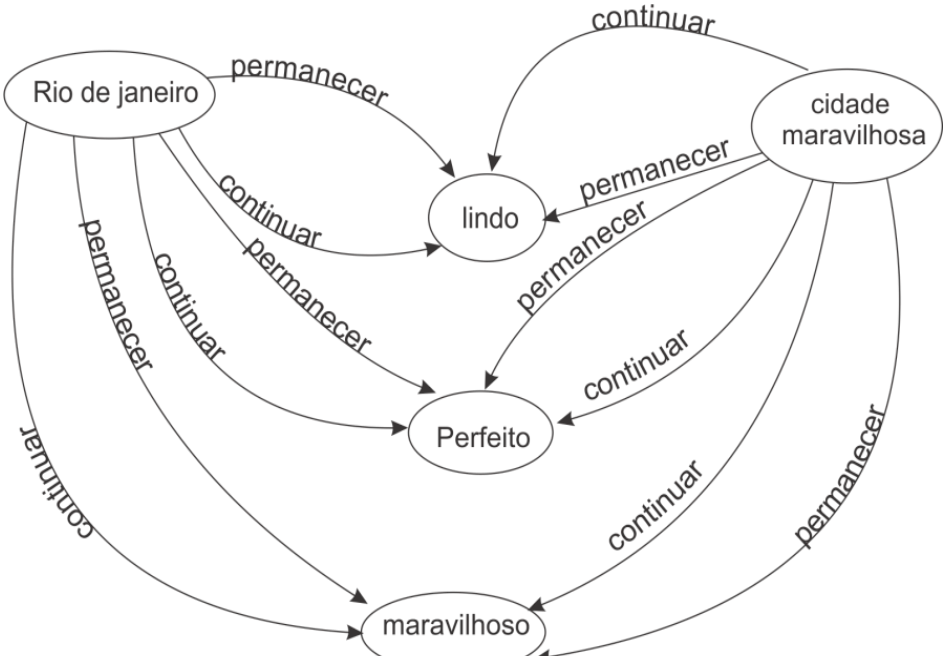


Figura 4. 30. Exemplo de concepts depois de finalizar a expansão de uma mr com conhecimento cultural.

A Figura 4.30 exemplifica graficamente *concepts* = *mr* $\cup$ α finalizando o processo de expansão conhecimento usando a base de conhecimento cultural como um banco de sinônimo cultural.

Ela mostra um conjunto de novas *mr* que unidas representam um único assunto, mas de forma diversificada que explora as possíveis variações da língua, ou seja, o conhecimento cultural das pessoas. Percebe-se que todas as *mr* possuem o mesmo significado podendo uma substituir a outra.

A próxima seção apresenta a evolução do método usando o processo de expansão de conhecimento apresentado nesta seção.

4.5.2.4 Evolução do método

O objetivo da evolução do método é igual ao anterior, ou seja, construir um conjunto de *mr* capaz de representar conhecimento sobre determinado assunto e, posteriormente, identificar usuários de SNS, através de comparações semânticas textuais, que se expressam de acordo com tal conhecimento, enfim, que estão falando sobre o mesmo assunto.

A diferença é que a representação de conhecimento, dessa vez, considera as variações no vocabulário das pessoas, podendo identificar pessoas que se expressam de forma diferente, pois é considerada a cultura do indivíduo através do seu vocabulário.

Inicialmente o algoritmo necessita de apenas uma oração que representa determinado assunto. Para exemplificar é usada a seguinte oração: *"Rio de Janeiro continua lindo"* (área 1 da Figura 4.30). A partir dela é construído uma *mr* (seção 4.4.2.3), nesse caso *continuar (Rio de janeiro, lindo)* (área 1 da Figura 4.31), que em seguida é expandida usando a base do OMCS-Br como fonte de sinônimo cultural (área 2 da Figura 4.31), como explicado na subseção 4.5.2.3. Finalmente, tem-se um conjunto α de novas outras *mr* que unido com a *mr* oriunda da oração forma o conjunto *concepts* (área 2 da Figura 4.31). *Concepts* é usado para identificar pessoas em SNSs que estão falando sobre o mesmo assunto (área 3 da Figura 4.31).

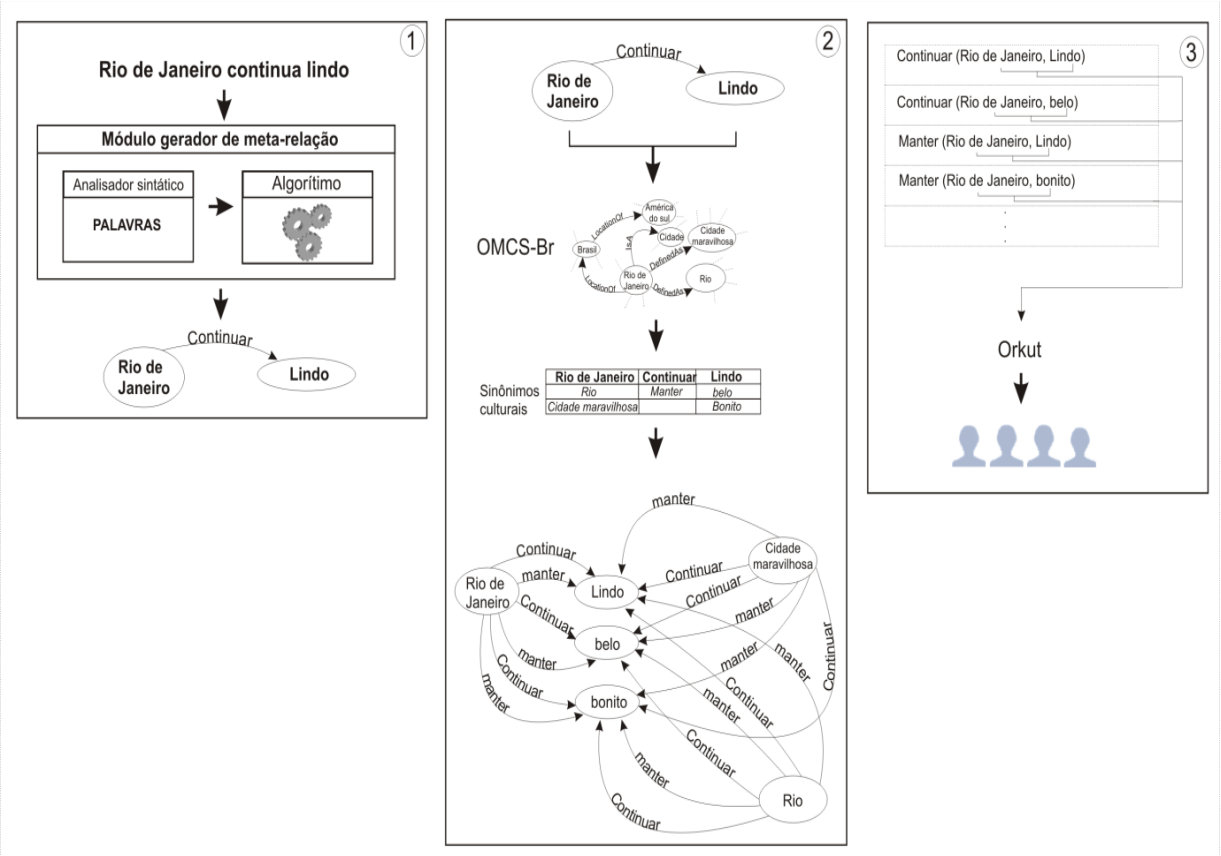


Figura 4. 31. Arquitetura do método que identifica pessoas que falam sobre o mesmo assunto em SNSs.

A busca por usuários em SNS é igual a versão anterior, ou seja, são usados pares de palavras retirados de cada uma das *mr* em busca de postagens nas Comunidades do Orkut. O processo de armazenamento das postagens recuperadas, bem como a verificação da existência da *mr* em questão na postagem e a recuperação do link do usuário também é idêntico a versão anterior do método.

A real mudança que essa iteração fez no método é o uso da base de conhecimento cultural como fonte de sinônimo. Dessa forma, há um uso mais efetivo do

conhecimento cultural armazenado pela base, o que faz com que o problema relacionado a esse ponto seja resolvido.

Além disso, toda a estrutura do método se tornou mais simples de ser mantida, isso consequentemente melhora o desempenho do algoritmo usado para implementá-lo.

4.5.3 Teste

O experimento dessa versão do método segue praticamente os mesmos passos da versão anterior. A diferença é que dessa vez, de acordo com a exigência da nova versão do método, é usada apenas uma oração. Os passos adotados são:

- Escolha de uma oração que representa um assunto;
- Representação de *concepts* para a busca por pessoas em SNSs;
- Uso do aplicativo da versão anterior que busca postagens no Orkut e as compara com as *mr* de *concepts*;

A oração usada como semente para representar o assunto usado como semente foi extraída de uma manchete de um portal de notícias on-line: *"Rio de Janeiro continua lindo?"*[19]. O conjunto *concepts* chegou a marca de 119 *mr* geradas pelo método. Na Tabela 4.19 são mostradas algumas das *mr* conseguidas.

Tabela 4. 19. Exemplo de mr geradas a partir de uma oração, considerando a expansão semântica.

Oração que representa o assunto	Exemplo do conteúdo de *concepts*
Rio de Janeiro continua lindo?	continuar (Rio de Janeiro, lindo); continuar (cidade maravilhosa, lindo); prosseguir (Rio, lindo); seguir (cidade maravilhosa, lindo);

Para usar as *mr* na busca por postagens, foi usada basicamente a mesma estratégia da versão anterior do método, ou seja, foi desenvolvido um aplicativo, baseado em um analisador de HTML, que vasculha a estrutura das páginas dos fóruns de discussão das Comunidades do Orkut. Cada postagem potencial encontrada, ou seja, que possui *s* e *c* da *mr* usada na busca, são armazenadas em um banco de dados associada a *mr* em questão (vide 4.4.2.5). Os resultados do experimento serão descritos na próxima seção.

4.5.3.1 Resultados

Foram identificadas 31 postagens que tinham pelo menos uma *mr* igual a uma das *mr* que representam o conhecimento usado como parâmetro na busca.

[19] http://www.monitormercantil.com.br/mostranoticia.php?id=72935.

A Tabela 4.20 mostra um ranking entre as *mr* que conseguiram os melhores resultados na busca. A *mr continuar (Rio de Janeiro, lindo)*, a melhor colocada, é a *mr* gerada a partir da oração usada como semente.

Tabela 4. 20. Ranking entre as *mr* com melhores resultados durante as buscas nas comunidades do Orkut.

mr	**Quantidade de postagens**	**(%)**
Continuar (Rio de Janeiro, lindo)	11	35,48
Continuar (Rio de Janeiro, belo)	4	12.9
Continuar (Rio, belo)	4	12.9
Continuar (Rio, maravilhoso)	4	12.9
Continuar (Cidade maravilhosa, lindo)	2	6.45
Continuar (Rio, lindo)	2	6.45

Na Figura 4.32 é apresentada uma relação entre a quantidade de postagens recuperadas das Comunidades do Orkut usando a *mr* oriunda da oração usada como semente, e as *mr* geradas com a expansão semântica. Observa-se que as *mr* geradas a partir da expansão semântica tiveram resultados mais satisfatórios.

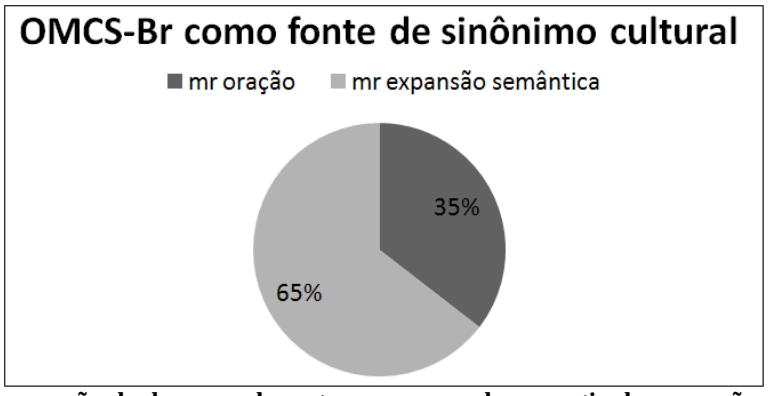


Figura 4. 32. Comparação do desempenho entre as *mr* geradas a partir da expansão semântica e a mr gerada a partir da oração.

Finalmente o conjunto β conseguiu um total de 17 usuários que estavam falando sobre o mesmo assunto, sendo 6 recuperados pela *mr continuar (Rio de Janeiro, lindo)*. O número de usuários agrupados não reflete o número de postagens recuperadas pelo método porque muitos dos usuários repetem o trecho de texto, que depois de ser transformado em uma *mr* é igual semanticamente a *mr* usada na busca, em outras postagens.

4.5.4 Apontar falhas

O uso da base de conhecimento cultural do Projeto OMCS-Br como uma base de sinônimos, mostrou ser a melhor alternativa para aproveitar o conhecimento cultural provido por ela. Os resultados na Figura 4.32 mostram esta confirmação.

Foram agrupados 11 usuários devido ao uso das *mr* expandidas semanticamente. Isso não seria possível em buscas que consideram apenas comparações textuais, e também não consideram a cultura das pessoas.

Aparentemente boa parte das postagens recuperadas com as *mr* geradas a partir da oração e, com o auxílio da base do OMCS-Br estavam fazendo referência ao assunto, isto é, a oração usada como semente. Isso quer dizer que os usuários do Orkut selecionados provavelmente estavam falando sobre o assunto em questão.

O problema relacionado ao falso-positivo percebido na versão anterior do método, aparentemente foi resolvido, pois com a versão atual é pouco provável que *mr* muito genéricas, como por exemplo, a *mr ser (Lula, presidente)* gerada na versão anterior do método, sejam geradas.

Com esses resultados conclui-se que, com a melhora significativa do método desde a primeira iteração, com essa iteração chega-se a versão final do método que objetiva identificar pessoas em SNSs que estão falando sobre o mesmo assunto, através de comparações semânticas textuais, no qual considera a cultura das pessoas no processo de busca.

4.6 Considerações finais

Neste capítulo foram descritas todas as iterações instanciadas com o uso da abordagem de trabalho proposta na seção 1.4, para se conseguir chegar ao método que identifica pessoas que estão falando sobre o mesmo assunto em SNSs, independente do vocabulário que elas utilizem, ou seja, da forma como elas se expressam.

A seção 4.5, a última iteração para obtenção do método, apresenta de fato a proposta deste trabalho, e que por isso, para observar a viabilidade de seu uso, foram realizados alguns estudos, descritos no próximo capítulo, com a participação de pessoas que disseram se os usuários dos SNSs selecionados pelo método estão falando sobre o assunto, usado como semente, através de suas postagens.

5 ESTUDO DE CASO

5.1 Considerações iniciais

Esse capítulo descreve o estudo de caso realizado com o intuito de observar o uso do método proposto por este trabalho, descrito integramente na seção 5.5, a fim de coletar a opinião das pessoas sobre o resultado obtido por meio do método.

Para realizar o estudo de caso foram definidas quatro etapas, cada uma delas se encontra em uma seção desse capítulo, como pode ser observado a seguir: seção 5.2 descreve sobre o planejamento do estudo de caso; seção 5.3 apresenta as etapas definidas para realizar o estudo de caso; seção 5.4 explica os questionários utilizados no estudo de caso; seção 5.5 a base teórica para a obtenção dos questionários; da seção 5.6 a 5.8 são apresentadas as instancias das etapas definidas na seção 5.3; e finalmente na seção 5.9 são apresentadas as considerações finais.

5.2 Planejamento do estudo de caso

Objetivo do estudo de caso: observar se o método proposto é capaz de identificar pessoas em SNSs que estão falando sobre o mesmo assunto, independente da forma como elas se expressam, como também, investigar se é valido o uso da base de conhecimento cultural do projeto OMCS-Br como uma base de sinônimos culturais.

Hipóteses:

(1) O método, proposto por este trabalho, identifica pessoas em SNSs que estão falando sobre o mesmo assunto, mesmo que utilizem vocabulários distintos;

(2) Conhecimento cultural pode ser usado para apoiar buscas textuais considerando diferenças culturais;

(3) Comparações semânticas textuais consideram o contexto de uma busca.

Método do estudo de caso: Para provar as hipóteses é preciso criar uma instância do método, por isso, algumas ações são necessárias:

Definir a semente, ou melhor, a oração usada como assunto para realizar a busca. É válido mencionar que há o cuidado de escolher a semente de forma a não sofrer qualquer influência pessoal.

Extrair da oração as relações semânticas a fim de representar computacionalmente a semântica da mesma, como também, identificar o assunto, contido na oração, para buscar usuários no SNS que estão falando sobre o mesmo.

Utilizar a base de conhecimento cultural como uma base de sinônimo cultural para expandir culturalmente as relações semânticas, com o intuito de aumentar a abrangência da busca no Orkut, pois, dessa forma, é possível considerar a diversidade do vocabulário dos usuários, ou seja, as diferenças culturais.

Para realizar essas ações:

(1) Foi utilizado o método desenvolvido pelo pesquisador do Laboratório de Interação Avançada do DC-UFSCar (LIA/DC-UFSCar);

(2) A oração usada, como semente, que dá origem ao assunto da busca foi recuperada de uma manchete de um portal de notícia on-line do Brasil;

(3) Foi convidado um grupo de pessoas para opinar sobre os resultados obtidos após o uso do método, com o intuito de observar a capacidade do mesmo de identificar pessoas que estão falando sobre o mesmo assunto no Orkut.

(4) Os dados colhidos desse estudo de caso foram coletados através de questionários, que foram entregues às pessoas convidadas.

***Seleção e Perfil dos Participantes*:** Para esse estudo de caso foi definido um grupo de convidados para observar a potencialidade do método. Houve o cuidado de definir o grupo de pessoas apenas por alfabetizados plenos (INSTITUTO PAULO MONTENEGRO, 2009), pois essas pessoas deveriam ter a capacidade de ler e interpretar um texto, para terem condições de identificar se uma postagem do Orkut, selecionada pelo método, está relacionada ao mesmo assunto que a oração usada como semente, mesmo que a postagem e a oração estivessem escritas de forma diferente.

Os alfabetizados plenos formam um grupo de pessoas cujas habilidades não mais impõem restrições de ler e compreender textos em situações usuais. Esse grupo, que deve ter pelo menos oito anos de escolaridade (oitava série), tem a capacidade de ler textos longos, orientar-se por subtítulos, localizar mais de uma informação em um texto, relacionar partes de um texto, comparar dois textos, distinguir fatos de opiniões, realizar inferências e sínteses (INSTITUTO PAULO MONTENEGRO, 2009).

5.3 Etapas do estudo de caso

Esse estudo foi divido em três etapas. Cada etapa possui um objetivo distinto descrito a seguir:

- Primeira: aplicar o método no Orkut para minerar postagens que estão relacionadas ao assunto expresso pela oração usada como semente;
- Segunda: explicar o processo do estudo de caso para as pessoas envolvidas e aplicar os questionários;
- Terceira: organizar e quantificar os resultados.

Na próxima seção são apresentados os questionários utilizados no estudo de caso, e em seguida as três etapas são descritas.

5.4 Questionários utilizados no estudo de caso

Nessa seção são apresentados os questionários elaborados para os participantes. Ressalta-se que as figuras a seguir ilustram apenas parte dos questionários, contudo, nos apêndices informados durante o texto há os questionários completos.

5.4.1 Questionário usado para coletar o perfil dos participantes

Esse questionário (APÊNDICE B), aplicado no fim da segunda etapa do estudo de caso, tem como objetivo coletar o perfil das pessoas. Sendo assim, identificar a sua faixa etária e o nível de escolaridade (Figura 5.1).

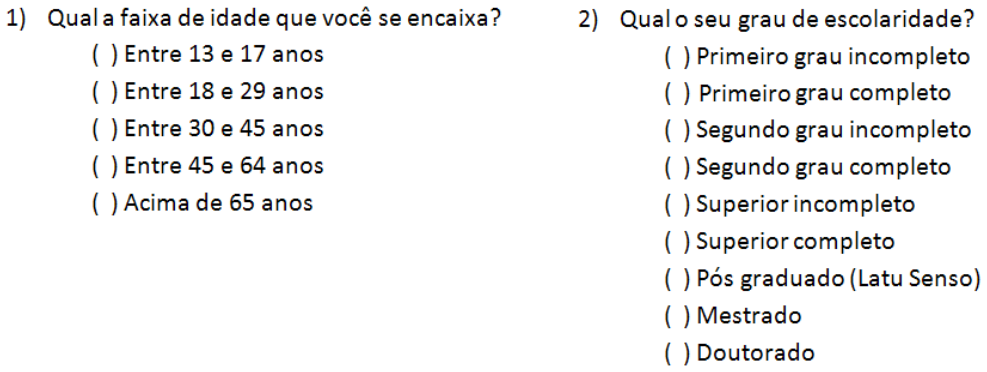


Figura 5. 1. Exemplos de perguntas do questionário sobre o perfil do participante

Através desse questionário pretende-se também identificar a familiaridade das pessoas com a leitura. Para isso foram elaboradas algumas perguntas que objetivam identificar se a pessoa lê e qual a frequência (Figura 5.2).

3) Você gosta de ler livros, revistas, jornais, ou outros?
   ( ) Sim
   ( ) Não
   Caso queira especificar que tipo de leitura gosta, escreva aqui:

4) Qual a sua frequência de leitura?
   ( ) Uma vez por dia
   ( ) Uma vez por semana
   ( ) Uma vez ao mês
   ( ) Outros – Especifique aqui:

Figura 5. 2. Perguntas do questionário para identificar a familiaridade dos participantes com a leitura.

Essas questões foram feitas porque a frequência de leitura do participante pode influenciar diretamente nos resultados do estudo de caso, pois, mesmo que o participante seja alfabetizado pleno, a dificuldade de leitura, como também, o não gosto por essa atividade, pode comprometer a compreensão da postagem, prejudicando a comparação da postagem obtida do Orkut com a oração utilizada como semente. Por meio dessas questões observou-se a possibilidade de verificar se a familiaridade do participante em relação a leitura pode influenciar no resultado.

No final da terceira etapa foi aplicado o questionário (APÊNDICE C) que objetiva observar a qualidade do método. As perguntas permitem aos participantes expressarem as suas opiniões sobre, se as postagens, conseguidas com a execução do método, estão relacionadas ao assunto usado para a busca; se a oração (assunto), usada como parâmetro, possui similaridade semântica com o trecho de texto da postagem obtido pelo Orkut. Com isso, espera-se coletar indícios de que os usuários estão ou não falando sobre o assunto em questão.

Existem outras perguntas que têm o objetivo de observar se o uso do conhecimento cultural da base do OMCS-Br pode apoiar buscas textuais considerando diferenças culturais. Finalmente, existe uma pergunta que objetiva identificar uma aplicabilidade para o método. Na Figura 5.3 é apresentado um exemplo do questionário. É válido mencionar que o objetivo de cada pergunta é descrito na seção 5.7.

Oração usada como assunto: “Lula defende José Sarney e diz que denúncias não tem fim”				
Textos ⬇	Você concorda que a pessoa que escreveu o texto à esquerda está falando sobre o mesmo assunto da oração (em vermelho acima)?	Considerando somente o trecho em destaque no texto à esquerda, você acha que ele possui o mesmo significado que a oração (em vermelho acima) usada como assunto?	Você acha que a pessoa que escreveu o texto à esquerda tem interesse sobre o assunto da oração (oração em vermelho acima)?	Se o assunto da oração (em vermelho acima) fosse de seu interesse, e se você e a pessoa que escreveu o texto fossem usuários de sites de redes sociais, você estaria disposto(a) a iniciar um contato com ele(a) ? Por quê?
Presidente defende Sarney e reclama de denúncias Um dia depois do discurso de José Sarney sobre a crise de ética que atinge o Senado (presidido por ele), o presidente Luiz Inácio Lula da Silva, em viagem pela Ásia, criticou a sequência de denúncias de irregularidades. "Elas não têm fim, e depois não acontece nada", disse Lula. Sarney, presidente do Senado, usou a tribuna da Casa ontem (16) para se defender das acusações de usar atos secretos para nomear parentes seus.	Concordo fortemente ( )	Possui muito ( )	Tem muito ( )	Muito disposto ( )
	Concordo ( )	Possui ( )	Tem ( )	Disposto ( )
	Concordo pouco ( )	Possui pouco ( )	Tem pouco ( )	Pouco disposto ( )
	Discordo ( )	Não possui ( )	Não tem ( )	Não disposto ( )
	Discordo fortemente ( )	Não tem nada a ver ( )	Não tem nada a ver ( )	De modo nenhum ( )
	Não posso opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( ) Porque:

Figura 5. 3. Exemplo do questionário usado para observar o uso método.

5.5 Base para a elaboração dos questionários

No questionário (APÊNDICE C) há apenas perguntas com alternativas. Essas perguntas têm como base o QUIS (*Questionnaire for User Interaction Satisfaction*) proposto por pesquisadores do Laboratório de Interação Humano-Computador (em inglês *Human-Computer Interaction*) da Universidade de Maryland (Estados Unidos da América), para avaliar subjetivamente a satisfação dos usuários em relação à interface humano-computador (CHIN *et al.*, 1998).

No questionário QUIS os graus de satisfação tem uma variância entre o grau mais baixo para o grau mais alto. De acordo com Chin *et al.*, (1998) o grau pode variar de 9 à 3. Por exemplo, Concordo Fortemente |__|__|__|__|__|__| Discordo fortemente. Para este caso, há 6 graus e apenas um espaço pode ser assinalado para representar o (grau) do quanto se está de acordo ou não com algo (PREECE *et al.*, 2002).

Como o método, proposto por este trabalho, trabalha com comparação semânticas textuais optou-se por adotar, em todos os questionários de avaliação do mesmo, a Escala de Likert (EVANS, 2008; WAINER, 2007) para representar os graus, uma vez que os mesmos são representados por pequenas sentenças, o que poderia facilitar a compreensão. Para o exemplo anterior a escala de Likert seria:

( ) Concordo Fortemente

( ) Concordo

( ) Ok

( ) Discordo

( ) Discordo Fortemente

( ) Não tenho como opinar

5.6 Primeira etapa do estudo de caso

Na primeira etapa do estudo de caso tem-se como objetivo a execução do método proposto por este trabalho, a fim de identificar pessoas que estão falando sobre o mesmo assunto nas comunidades do Orkut.

O primeiro passo é identificar apenas uma oração como semente, que represente um assunto, para iniciar a busca, entretanto, com o intuito de observar o método com mais detalhes optou-se por realizar a busca considerando três assuntos distintos ao invés de apenas um assunto. Na Tabela 5.1 são apresentas as orações escolhidas e as respectivas URLs dos portais de notícias de onde elas foram extraídas.

Tabela 5. 1. Orações usadas como assunto neste estudo de caso.

Orações	URLs onde estão as manchetes
"O Rio de Janeiro continua lindo?"	http://super.abril.com.br/cotidiano/rio-janeiro-continua-lindo-447914.shtml
"Michael Jackson morre aos 50 anos."	http://oglobo.globo.com/cultura/mat/2009/06/25/michael-jackson-morre-aos-50-anos-756516758.asp
"Lula defende José Sarney e diz que denúncias não têm fim."	http://www1.folha.uol.com.br/folha/brasil/ult96u582293.shtml

A primeira oração representa um assunto relacionado a turismo. Nesse contexto, pretende-se identificar pessoas que estão falando sobre o "Rio de Janeiro" e, que a considera uma cidade linda, com isso, há a possibilidade de dizer que essas pessoas têm chances de se interessarem por esse assunto, por terem a mesma opinião sobre ele.

A segunda oração representa um assunto relacionado a celebridades. Com ela pretende-se identificar usuários do Orkut que estão falando sobre a morte de "Michael Jackson", que ocorreu no ano de 2009.

Finalmente, a terceira representa o assunto política brasileira. O uso dessa oração objetiva identificar usuários do Orkut que estão falando sobre a suposta proteção do presidente "Lula" ao senador "José Sarney".

5.6.1 Resultados dessa etapa

Com a primeira oração "O Rio de janeiro continua lindo" foi possível gerar 119 meta-relações. Alguns exemplos são: *continuar (Rio de Janeiro, lindo), permanecer (Rio, belo),* etc.

Foram identificadas 31 postagens que continham no mínimo uma meta-relação. Algumas postagens são:

1. *"**O Rio de Janeiro continua lindo**, sim. Os problemas do RJ são os mesmos de qq cidade grande de um país subdesenvolvido ou em desenvolvimento, como preferem alguns";*
2. *"Estou há quase 11 anos morando em Cascavel, no Paraná, ía ao Rio uma vez por ano, estava muito desacostumada, tinha medo da violência, neste ano, passei 5 meses na cidade, e vi que ñ é nada disso, o **Rio continua maravilhoso** como sempre foi. Eu quero voltar!!".*

É possível observar, ao ler as duas postagens, que o trecho de texto em destaque na segunda postagem pode estar relacionado com o trecho em destaque da primeira postagem "Rio de janeiro continua lindo", mesmo estando escrito de forma diferente. Isso é um indício de que o método proposto por este trabalho consegue identificar pessoas que estão falando sobre o mesmo assunto, independente do vocabulário usado.

Com esse indício, percebe-se a possibilidade de satisfazer uma das hipóteses levantadas no inicio desse estudo de caso, em que a base de conhecimento cultural do OMCS-Br pode ser usada para apoiar buscas textuais considerando diferenças culturais, além disso, mostra que o método identifica pessoas que estão falando sobre o mesmo assunto, mesmo que elas usem vocabulários distintos. Na Tabela 5.2 há alguns exemplos de trechos de postagens identificados pelo método com a respectiva meta-relações usada.

Tabela 5. 2. Exemplo de trechos de postagens identificadas com o uso do primeiro assunto.

Meta-relação	Trecho das postagens que possibilitou a identificação do usuário.
Continuar (Rio, belo)	*...lindo. O **Rio continua belo**. O Rio...*
Continuar (Rio, maravilhoso)	*...nada disso, **o Rio continua maravilhoso** como sempre...*
Continuar (Cidade Maravilhosa, linda)	*...**cidade maravilhosa continua linda**, como sempre, mas...*
Continuar (Rio, lindo)	***... Rio continua lindo** com grandes...*

Com a segunda oração "Michael Jackson morre aos 50 anos" foi possível gerar 32 meta-relações. Alguns exemplos são: *morrer (Michael Jackson, 50 anos), morrer (Rei do pop, 50 anos), bater as botas (Rei do pop, 50 anos),* etc. Essas meta-relações possibilitaram encontrar 22 postagens.

Alguns trechos de postagens relacionados a esse assunto podem ser observados na Tabela 5.3. Nessa Tabela, assim como na Tabela 5.2, há tanto os trechos quanto as respectivas metas-relações usadas na busca.

Tabela 5. 3. Exemplo de trechos de postagens identificadas com o uso do segundo assunto.

Meta-relação	Trecho das postagens que possibilitou a identificação do usuário.
morrer (Michael Jackson, 50 anos)	***...Michael Jackson morre aos 50 anos*** *após parada card...*
falecer (Michael Jackson, 50 anos)	***...Michael Jackson faleceu aos 50 anos*** *em Los Angeles, Califórnia...*
falecer (cantor, 50 anos)	*...José Lewgoy. O* ***cantor faleceu aos 50 anos****, em 14 de fevereiro de 1996...*
morrer (Rei do pop, 50 anos)	*..."O* ***Rei do pop morre aos 50 anos****'' .isso é o q diz o vídeo...*

Já com a terceira oração "*Lula defende José Sarney e diz que denúncias não têm fim*" conseguiu-se gerar 70 meta-relações, como por exemplo, *proteger (Lula, José Sarney)* e *falar (Presidente, denúncias não têm fim).* Alguns trechos obtidos por meio dessas meta-relações estão na Tabela 5.4.

Tabela 5. 4. Exemplo de trechos de postagens identificadas com o uso do terceiro assunto.

Meta-relação	Trecho das postagens que possibilitou a identificação do usuário.
Dizer (Lula, denúncias não tem fim)	***...Lula*** *defende José Sarney e* ***diz que denúncias não têm fim.*** *da Agência Brasil...*
Defender (Presidente, Sarney)	*...****Presidente defendendo Sarney****. Petistas comentem...*
apoiar (lula, Sarney)	***...Lula apoia Sarney*** *de olho na ABL...*
proteger (lula, ladrão)	*...fuma o* ***lula protege ladrão*** *por isso ele...*

Após observar os trechos obtidos, foi possível perceber uma possível falha do método, por exemplo, no terceiro trecho da Tabela 5.3, o método identificou uma postagem onde o usuário diz: *"cantor faleceu aos 50 anos"*, no entanto, nesse caso não está claro na meta-relação de qual cantor se trata, assim durante a busca podem ocorrer postagens que não estão relacionadas a morte de Michael Jackson, mas sim, nesse caso, o falecimento de qualquer cantor aos 50 anos.

Essa falha percebida é uma das justificativas para a realização de um estudo de caso, pois a proposta é permitir que os participantes ao lerem as postagens recuperadas possam dizer se a mesma está relacionada ao assunto em questão.

Na próxima seção é apresentado como foi conduzido o estudo de caso com os participantes e, em seguida há o resultado desse estudo.

5.7 Segunda etapa do estudo de caso

Nessa etapa o objetivo foi explicar aos participantes o estudo de caso, bem como a forma com que eles poderiam observar o resultado do método e, expressar as suas opiniões por meio dos questionários. Durante a explicação alguns assuntos foram abordados, tais como: o objetivo do estudo e os questionários a serem utilizados.

Os participantes receberam dois questionários. O primeiro, como descrito anteriormente, tem o objetivo de colher o perfil do participante e o segundo de avaliação do método. No segundo questionário cada participante recebeu uma tabela com a oração, usada como assunto na busca, e postagens, relacionadas a cada oração recuperadas pelo método, que representam o que alguns dos usuários do Orkut escreveram sobre o assunto.

Cada participante foi chamado a ler as orações e as postagens e a responder quatro questões, considerando cada postagem individualmente. Por exemplo, na Figura 5.4 é mostrado como foi aplicado o questionário ao participante. Na área (A) é mostrada a oração usada como assunto na busca. Na área (B) é apresentada a postagem. Na área (C), o participante deve preencher uma alternativa que representa a sua opinião considerando a pergunta feita.

Ⓐ **Oração usada como assunto: "Lula defende José Sarney e diz que denúncias não tem fim"**				
Textos ⬇	Você concorda que a pessoa que escreveu o texto à esquerda está falando sobre o mesmo assunto da oração (em vermelho acima)? Ⓒ	Considerando somente o trecho em destaque no texto à esquerda, você acha que ele possui o mesmo significado que a oração (em vermelho acima) usada como assunto? Ⓒ	Você acha que a pessoa que escreveu o texto à esquerda tem interesse sobre o assunto da oração (oração em vermelho acima)? Ⓒ	Se o assunto da oração (em vermelho acima) fosse de seu interesse, e se você e a pessoa que escreveu o texto fossem usuários de sites de redes sociais, você estaria disposto(a) a iniciar um contato com ele(a) ? Por quê? Ⓒ
Presidente defende Sarney e reclama de denúncias Um dia depois do discurso de José Sarney sobre a crise de ética que atinge o Senado (presidido por ele), o presidente Luiz Inácio Lula da Silva, em viagem pela Ásia, criticou a sequência de denúncias de irregularidades. "Elas não têm fim, e depois não acontece nada", disse Lula. Sarney, presidente do Senado, usou a tribuna da Casa ontem (16) para se defender das acusações de usar atos secretos para nomear parentes seus. Ⓑ	Concordo fortemente ( )	Possui muito ( )	Tem muito ( )	Muito disposto ( )
	Concordo ( )	Possui ( )	Tem ( )	Disposto ( )
	Concordo pouco ( )	Possui pouco ( )	Tem pouco ( )	Pouco disposto ( )
	Discordo ( )	Não possui ( )	Não tem ( )	Não disposto ( )
	Discordo fortemente ( )	Não tem nada a ver ( )	Não tem nada a ver ( )	De modo nenhum ( )
	Não posso opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( ) Porque:

Figura 5. 4. Exemplo de como o questionário foi aplicado ao participante do estudo de caso.

A pergunta número (1) tem o objetivo de observar se a pessoa que escreveu a postagem está falando sobre o assunto expresso pela oração usada como assunto para busca. Essa pergunta tem como intuito coletar indícios de que o objetivo deste trabalho foi ou não alcançado.

A pergunta número (2) objetiva comparar e identificar se o trecho de texto, em destaque na postagem, tem similaridade semântica com a oração usada como parâmetro. Essa pergunta foi elaborada para observar se o método é capaz de realizar buscas usando comparações semânticas textuais, como proposto, como também, se o conhecimento cultural, provido pela base do OMCS-Br, pode apoiar buscas textuais considerando diferenças culturais.

Com a comparação entre as respostas dessa pergunta (2) com a pergunta (1), será possível verificar se comparações semânticas textuais podem ou não ter bons resultados quando é usada para buscar pessoas que estão falando sobre um mesmo assunto em SNSs.

A pergunta número (3) e (4) são um tanto subjetivas. A pergunta (3) busca coletar a opinião dos participantes a respeito do quanto o usuário responsável pela postagem demonstraria ter interesse pelo assunto usado na busca. A (4) objetiva identificar se o método pode apoiar sistemas de recomendação social.

Na próxima seção são apresentados os resultados obtidos com o estudo de caso.

5.8 Terceira etapa do estudo de caso

Nessa terceira etapa o objetivo foi apresentar os resultados obtidos com o estudo de caso. Os questionários foram entregues ou enviados por e-mail para 42 pessoas, mas, apenas 19 responderam.

Algo que pode ter inibido a participação dos outros participantes é o tamanho do questionário, pois para cada oração havia algumas postagens. Considerando os três assuntos usados no estudo de caso, somavam 38 postagem com 4 perguntas para cada. Abaixo segue a distribuição de postagens para cada assunto:

- "Rio de janeiro continua lindo" – 15 postagens;
- "Michael Jackson morre aos 50 anos" – 9 postagens;
- "Lula defende José Sarney e diz que denúncias não têm fim" – 14 postagens.

Vale ressaltar que o método conseguiu recuperar 77 postagens, no entanto, após a entrega do questionário aos primeiros participantes, foi percebido que o mesmo estava muito extenso e, por isso, para evitar a desistência dos participantes em respondê-lo foram eliminadas, de modo aleatório, 50% das postagens de cada assunto. Mesmo assim houve um

desinteresse considerável entre boa parte das pessoas convidadas, o que ocasionou em enviar o questionário para outras pessoas.

Na Figura 5.5 há um gráfico mostrando a distribuição de todas as postagens conseguidas pelos três assuntos considerados no estudo de caso. Por meio do gráfico é possível observar que 53% são postagens que foram recuperadas com o apoio de conhecimento cultural usado como sinônimo cultural, isso significa que apenas com a oração, sem a utilização da base cultural, foi possível identificar 47% das postagens e, com o uso da base cultural para expandir o vocabulário 53%.

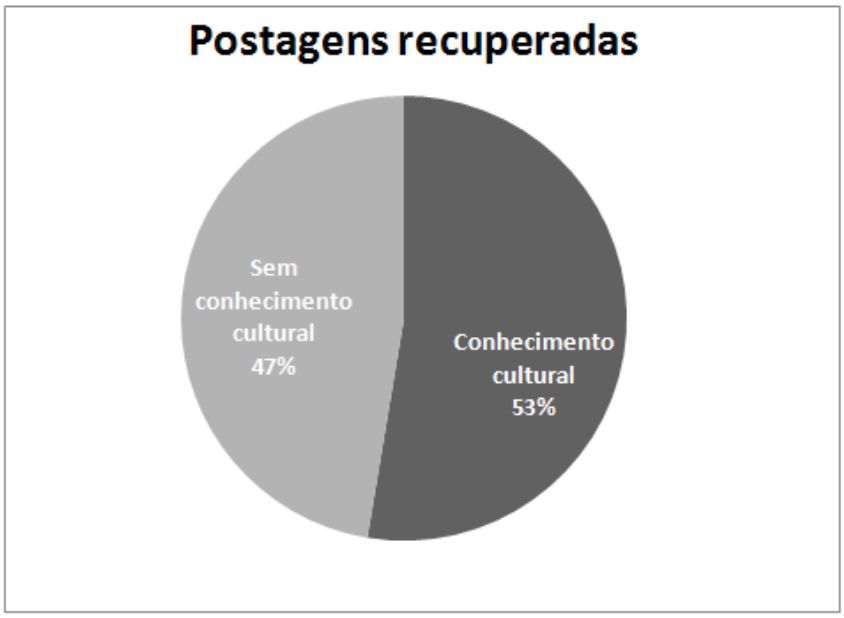


Figura 5. 5. Comparação entre a quantidade de postagens recuperadas com e sem o apoio de conhecimento cultural.

Por exemplo, a postagem: "*Mas o **Rio continua Maravilhoso** e é o único lugar que me desperta sempre a vontade de voltar e passar as minhas férias, apesar de ...*". Foi recuperada por causa do apoio da base de conhecimento cultural, pois a oração original usada como assunto é "*Rio de Janeiro continua lindo*".

Nesse caso "*Rio de janeiro*" foi substituído pelo sinônimo cultural "*Rio*", encontrado na busca na base como: *definedAs(Rio de Janeiro, Rio)*. Já "*lindo*" foi substituído por "*maravilhoso*" encontrado na base como: *definedAs(lindo, maravilhoso)*. O final do processo forma a oração encontrada na postagem, que no caso é "*Rio continua maravilhoso*".

Como descrito anteriormente, 47% das postagens foram recuperadas sem o uso de conhecimento cultural, ou seja, usando os componentes da oração original usada como assunto (Figura 5.5). Por exemplo, a postagem: "*Sim! **O Rio de Janeiro continua LINDO!** Mesmo que muitos torçam pelo contrário*", foi recuperada usando apenas a representação

semântica derivada da oração usada como assunto, isto é, a *mr continuar (Rio de Janeiro, lindo)*. Nas próximas subseções são mostradas as análises para cada uma das quatro perguntas usadas no estudo de caso com o intuito de permitir aos participantes expressarem suas opiniões após comparar as orações com as postagens obtidas.

5.8.1 Análise sobre a Pergunta 1

A Pergunta 1 do questionário foi apresentada da seguinte forma:

"*Você concorda que a pessoa que escreveu o texto à esquerda está falando sobre o mesmo assunto da oração (em vermelho acima)?*

( ) concordo fortemente

( ) concordo

( ) concordo pouco

( ) discordo

( ) discordo fortemente

( ) Não tenho como opinar".

O "*texto a esquerda*" se refere a postagem recuperada pelo método. A "*oração*" se refere a uma das 3 orações usadas como assunto para as buscas nas comunidades do Orkut.

Essa pergunta visa coletar indícios que mostrem se o objetivo principal do trabalho, que é identificar pessoas em SNS que estão falando sobre o mesmo assunto, foi alcançado, além disso, essa pergunta é usada para verificar se as hipóteses 1 e 2, levantadas pelo estudo de caso, foram alcançadas.

No gráfico da Figura 5.6 são apresentados os resultados totais referentes aos 3 assuntos usados na busca.

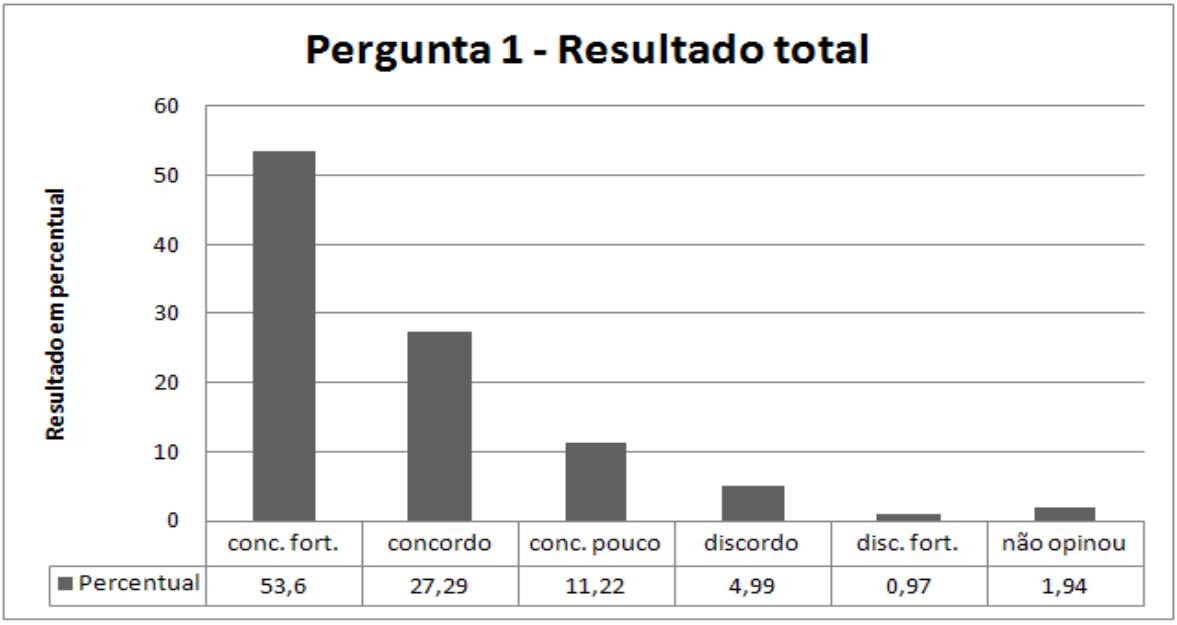


Figura 5. 6. Resultado referente às repostas da Pergunta 1 para os três assuntos usados no estudo de caso.

Se considerar apenas as duas primeiras alternativas (concordo fortemente e concordo) como as mais relevantes para identificar se as postagens estão relacionadas com a oração, ou seja, se foi possível identificar pessoas que estão falando sobre o mesmo assunto da oração, é possível mencionar que o método conseguiu 80,89% de sucesso entre as 38 postagens analisadas pelos participantes.

Esse número mostra, segundo a opinião dos participantes, que o método conseguiu atingir o objetivo, pois 80,89% é um número, considerado por este trabalho, de grande expressividade. Mesmo porque o número de erros não passa de 6%, isso considerando as duas últimas alternativas (discordo e discordo fortemente) como caso de insucesso. Se considerar as três primeiras alternativas (concordo fortemente, concordo e concordo pouco), é possível descrever que o método atingiu um percentual de acerto de 92,11%.

Na busca usando o método disponível no Orkut as possibilidades ficam muito limitadas. Por exemplo, usando a oração "*O Rio de Janeiro continua lindo*" entre aspas, que é a melhor forma de considerar o assunto; a busca tem praticamente a mesma eficiência que o método proposto por este trabalho sem o uso da base cultural.

A diferença é que na busca do Orkut são consideradas apenas os textos idênticos, ou seja, orações, como por exemplo, "*O Rio de Janeiro continuará lindo*", que semanticamente é similar a "*O Rio de Janeiro continua lindo*", não é considerada. Nesse caso, a variação temporal da oração é ignorada, algo que o método proposto por este trabalho consegue identificar e considerar.

Postagens que possuem orações escritas de forma diferente, como por exemplo, "*O Rio continua belo*", não são consideradas pelo mecanismo de busca do Orkut, no entanto, segundo 94,17 % dos participantes do estudo de caso, essa oração é semanticamente similar à oração usada na busca, por isso, o método identificou com sucesso uma postagem relacionada a oração. Esse tipo de comparação, que considera a cultura das pessoas, é o ponto principal do método proposto por este trabalho. Algumas outras comparações entre o método proposto por este trabalho e a busca disponível no Orkut estão na Tabela 5.5.

Tabela 5. 5. Comparação entre o método de busca disponível no Orkut com o método proposto por este trabalho.

A busca	Método do Orkut	Método proposto por este trabalho
Considera comparações textuais. Por exemplo: "*O Rio de Janeiro continua lindo*" = "*O Rio de Janeiro continua lindo*".	Sim	Sim
Considera comparações semânticas textuais. Por exemplo, "*Rio de Janeiro continua lindo*" = "*Rio de Janeiro permanece lindo*".	Não	Sim
Considera variação de tempo nas comparações. Por exemplo: "*Rio de Janeiro continua lindo*" = "*Rio de Janeiro continuará lindo*"	Não	Sim
Considera a cultura dos usuários. Por exemplo: "*Rio de Janeiro continua lindo*" = "*Cidade maravilhosa permanece bela*"	Não	Sim

É possível ter indícios, ao observar a Tabela 5.5, da vantagem que o método proposto por este trabalho possui ao considerar o contexto cultural e a comparação semântica entre as postagens e as orações utilizadas para realizar as buscas.

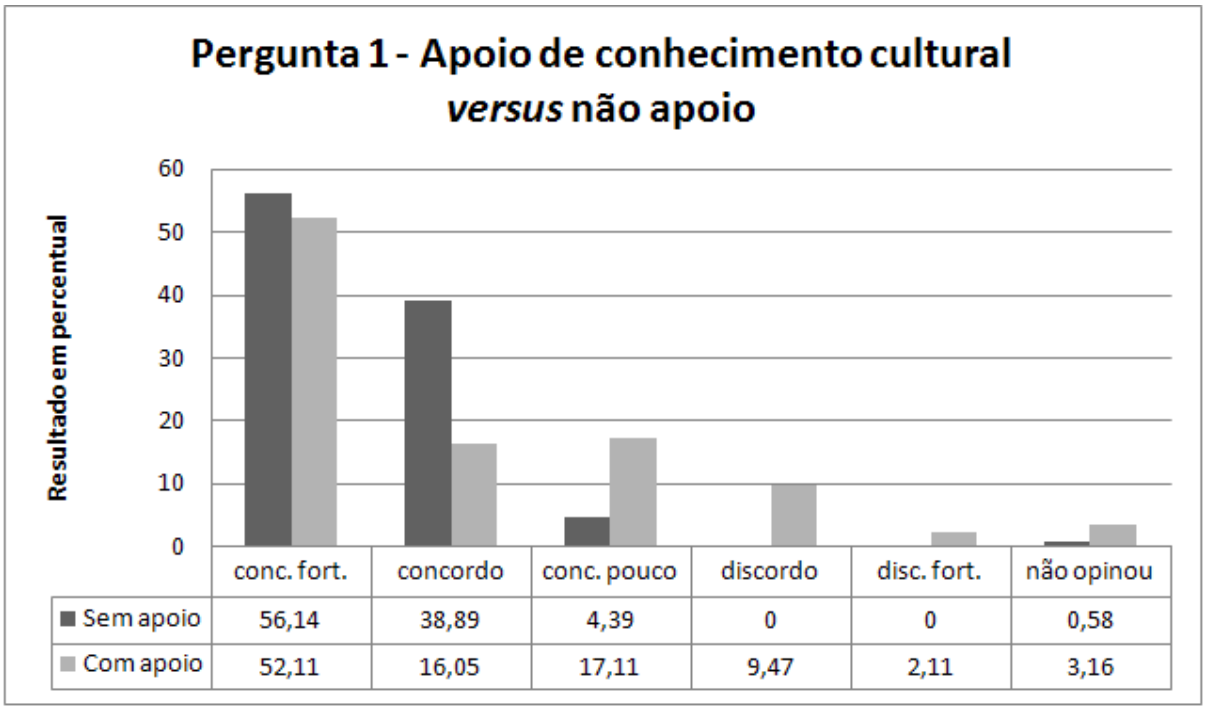


Figura 5. 7. Comparação entre o resultado do "uso" e o "não uso" de conhecimento cultural em relação às respostas da Pergunta 1.

No gráfico apresentado na Figura 5.7 há a comparação entre o resultado do "uso" e do "não uso" de conhecimento cultural na busca. No "não uso" são consideradas apenas postagens que tiveram as mesmas palavras das orações, como por exemplo, "*O Rio de Janeiro continua lindo*", "*O Rio de Janeiro continuará lindo*", "*O Rio de Janeiro continuou lindo*", etc. e, o resultado com o "uso" de conhecimento cultural na busca, são consideradas as postagens que possivelmente estão em um mesmo contexto, independente da forma com que são escritas, como por exemplo, "*Cidade maravilhosa continua linda*", "*Rio permanece belo*", etc.

Ao analisar o resultado dessa comparação, percebeu-se que o "uso" de conhecimento cultural leva desvantagem quanto à eficiência nos resultados, pois se considerar apenas as duas primeiras alternativas (concordo fortemente e concordo) como caso de sucesso, o "não uso" tem um acerto, segundo a opinião dos participantes do estudo de caso, de 95,47% contra 68,16% do "uso" de conhecimento cultural.

O resultado obtido com o "uso" de conhecimento cultural mostrou indícios de que o uso da base do OMCS-Br como banco de sinônimos culturais é uma possível solução de sucesso para identificar pessoas que falam dos mesmos assuntos, mesmo que se expressam de maneiras diferentes, uma vez que as pessoas identificadas falam sobre um determinado assunto em questão, mesmo tendo algumas diferenças culturais na forma de se expressarem.

Dessa forma, aqui a cultura é usada como um recurso importante para aproximar as pessoas e, não para distanciá-las, pois mesmo expressando de formas diferentes, há a possibilidade das pessoas terem interesses em comum.

Ao considerar as duas últimas alternativas da Pergunta 1 (discordo e discordo fortemente) como caso de insucesso do método, o "uso" de conhecimento cultural teve 8.69% de erro, isto é, identificou postagens em que os participantes disseram que as pessoas não estavam falando sobre os assuntos em questão. Já o "não uso" de conhecimento cultural teve 2.93% de erro.

A oração "*A praia continua linda*" é um exemplo de uma oração, identificada em uma postagem, que levou o método a cometer um desses erros. Nesse caso o problema foi percebido na *mr continuar (Praia, linda)* que foi utilizada para identificar uma postagem, que os participantes disseram que estava em outro contexto que não a praia do Rio de Janeiro. Essa *mr* foi gerada a partir da relação de Minsky *DefinedAs(Rio de Janeiro, praia)*, isto é, "*Rio de Janeiro*" foi substituído por "*praia*".

Percebe-se que a *mr* não deixa claro que o assunto que se procura é sobre "*Rio de janeiro continua lindo*". Apesar de "*Praia*" ser definida como sinônimo de "*Rio de Janeiro*" por um colaborador do projeto OMCS-Br, quando se usa esse conhecimento cultural é distorcido a semântica da oração, fazendo com que ela expresse algo indefinido em relação ao assunto: "*Praia continua linda*". Qual "*Praia*"?

Abaixo segue o exemplo da postagem que provocou o equívoco:

"Oi!! Idem, idem! Pleeease, alguém aqui é da época que andar de mobilete no Nove era tuuudo!? Quando de noite não tinha lâmpada na praia, de dia tinha siri e todo mundo podia entrar no iate prá nadar?Aiiii, saudade deliciosa!É bom saber que a praia continua linda e querida por muitos! Bjocas a todos!".

O erro percebido em relação ao "uso" de conhecimento cultural é difícil de ser controlado, pois o conhecimento capturado da base que define culturalmente o conceito procurado, como é o caso de "*Rio de Janeiro*" ser considerado sinônimo de "*Praia*", quando aplicado em um contexto maior, como de uma postagem de um usuário em que no caso deste trabalho pode acabar ficando sem sentido.

5.8.2 Análise sobre a Pergunta 2

A Pergunta 2, usada no questionário, é a seguinte:

"Considerando somente o trecho em destaque no texto à esquerda, você acha que ele possui o mesmo significado que a oração (em vermelho acima) usada como assunto?

( ) Possui muito

( ) Possui

( ) Possui pouco

( ) Não possui

( ) Não tem nada a ver

( ) Não tenho como opinar"

O "*texto a esquerda*" se refere à postagem recuperada pelo método. A "*oração*" se refere a uma das 3 orações usadas como assunto para as buscas. Uma novidade nessa Pergunta é o "*trecho de texto*". Ele é um trecho em destaque na postagem, como por exemplo, o trecho "*Rio de Janeiro continua lindo*" em destaque na postagem abaixo:

"*Sim!* ***O Rio de Janeiro continua LINDO****! Mesmo que muitos torçam pelo contrário.".*

Essa pergunta, direcionada apenas ao trecho de texto, foi usada com o objetivo de observar se as *mr* usadas como parâmetro nas buscas estão representando fielmente os assuntos em questão, além disso, verifica-se também se as *mr* derivadas a partir do uso da base de conhecimento cultural não perdem a representatividade do assunto.

Tendo um resultado positivo nessa pergunta, confirma-se novamente a segunda hipótese levantada por esse estudo de caso, além disso, espera-se observar se as comparações semânticas textuais, proposta por este trabalho, estão sendo realizada com êxito.

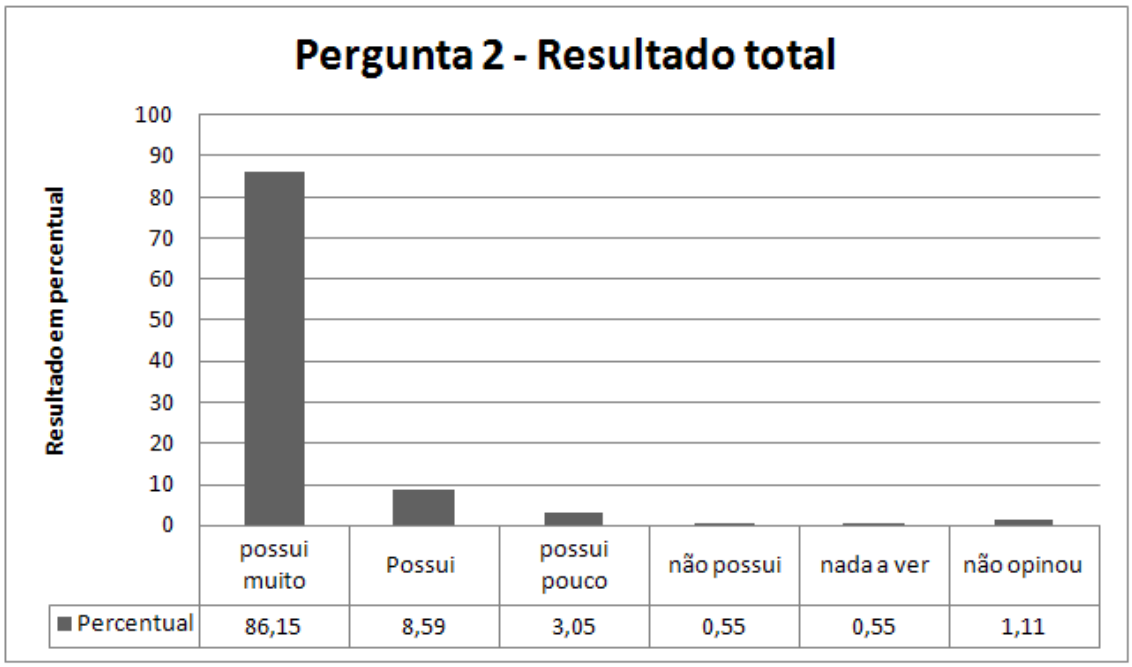


Figura 5. 8. Resultado referente às repostas da Pergunta 2 para os três assuntos usados no estudo de caso.

De modo geral o gráfico apresentado na Figura 5.8 mostra que o método obteve resultado satisfatório nas comparações semânticas textuais, pois se considerar as três primeiras alternativas (Possui muito, Possui e Possui pouco) da Pergunta 2 como uma referência para a o sucesso, o método consegui 97,78% de acerto.

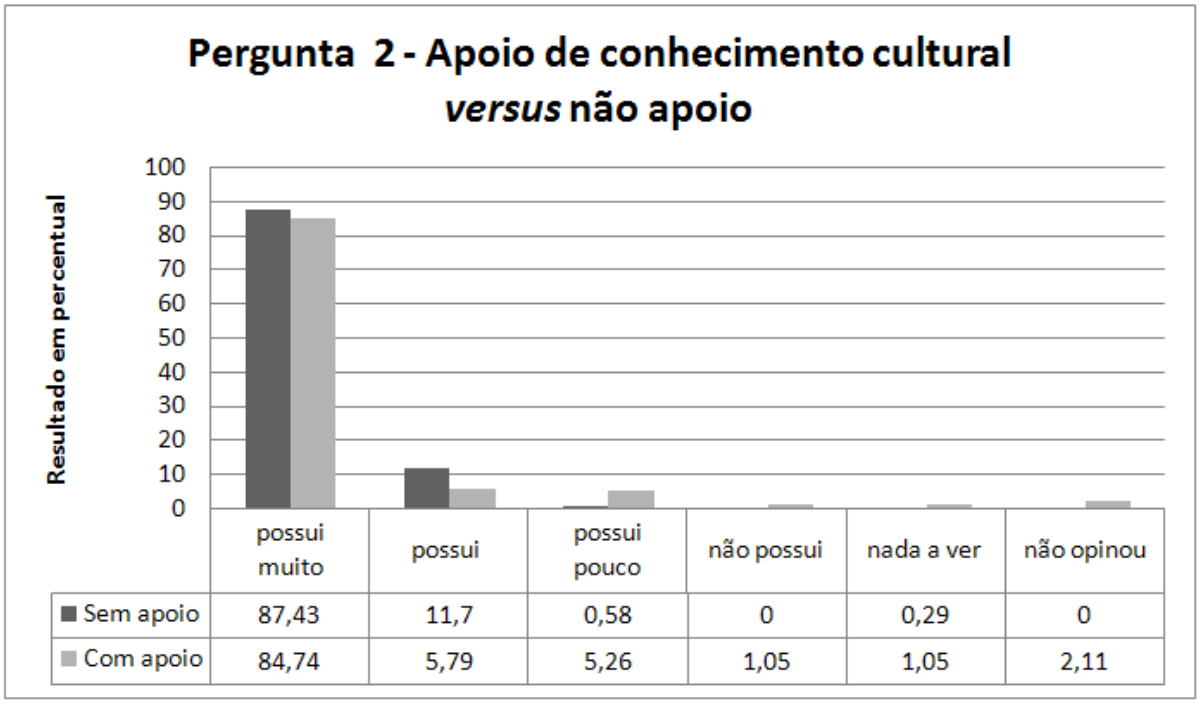


Figura 5. 9. Comparação entre o resultado do "uso" e o "não uso" de conhecimento cultural em relação às respostas da Pergunta 2.

Na Figura 5.9 é apresentada uma comparação entre o "uso" de conhecimento cultural na busca e o "não uso". Observa-se que considerando as três primeiras alternativas

(possui muito, possui e possui pouco) da Pergunta 2, houve 90.57% de sucesso com o "uso" de conhecimento cultural. Já o não uso, mostrou-se mais eficaz, com 99.13% de acerto.

Essa diferença entre os dois dados é explicada pelo mesmo fato percebido na análise dos resultados obtidos pela Pergunta 1. O erro pode ocorrer, pois o conhecimento cultural extraído da base é usado em um contexto maior, já o "não uso" busca apenas por palavras idênticas o que facilita encontrar sentenças também idênticas. No caso da análise feita com as respostas da Pergunta 1 houve um erro de 8.69%, nesse caso o erro foi de 2.10%.

Essa diferença entre os dois casos de erros é porque na Pergunta 1 os participantes do estudo de caso consideravam a postagem como um todo, inclusive seu contexto. No caso da Pergunta 2, os participantes consideraram apenas o trecho de texto em destaque, por isso, que possivelmente se observou mais erros na análise da Pergunta 1.

Com esses resultados este trabalho considera que há indícios de que a segunda hipótese levantada por esse estudo de caso é satisfeita, além disso, foi percebido que as buscas por comparações semânticas, proposta por este trabalho, estão sendo realizadas com uma considerável precisão.

5.8.3 Análise sobre a Pergunta 3

A Pergunta 3:

"*Você acha que a pessoa que escreveu o texto à esquerda tem interesse sobre o assunto da oração (oração em vermelho acima)?*

( ) Tem muito

( ) Tem

( ) Tem pouco

( ) Não tem

( ) Não tem nada a ver

( ) Não tenho como opinar".

Essa pergunta é um tanto subjetiva, que ultrapassa o escopo deste trabalho. Ela pretende identificar se as pessoas identificadas pelo método poderiam ter interesse pelos assuntos usados na busca. Esse resultado, como nos anteriores, se baseia na opinião dos participantes do estudo de caso.

É válido mencionar que o resultado da Pergunta 1 poderia ser utilizado para ter indícios de que um usuário do Orkut, responsável por uma determinada postagem, tem interesse por um determinado assunto, entretanto, este trabalho assume que uma pessoa quando fala sobre um assunto, possivelmente ela poderia ter interesse pelo mesmo. Por esse motivo foi aplicado essa pergunta aos participantes, para que eles pudessem observar mais do que apenas palavras existentes nas postagens, mas sim todo um contexto e interesse que pudesse estar implícito no texto, algo que seria difícil de ser feito computacionalmente.

Na Figura 5.10 são apresentados os dados dessa avaliação. Observa-se que os participantes do estudo de caso consideraram que em 77,84% das postagens os usuários demonstraram ter interesse pelo assunto em questão. Isso considerando como satisfatórias as duas primeiras alternativas da Pergunta 3 (Tem muito e Tem).

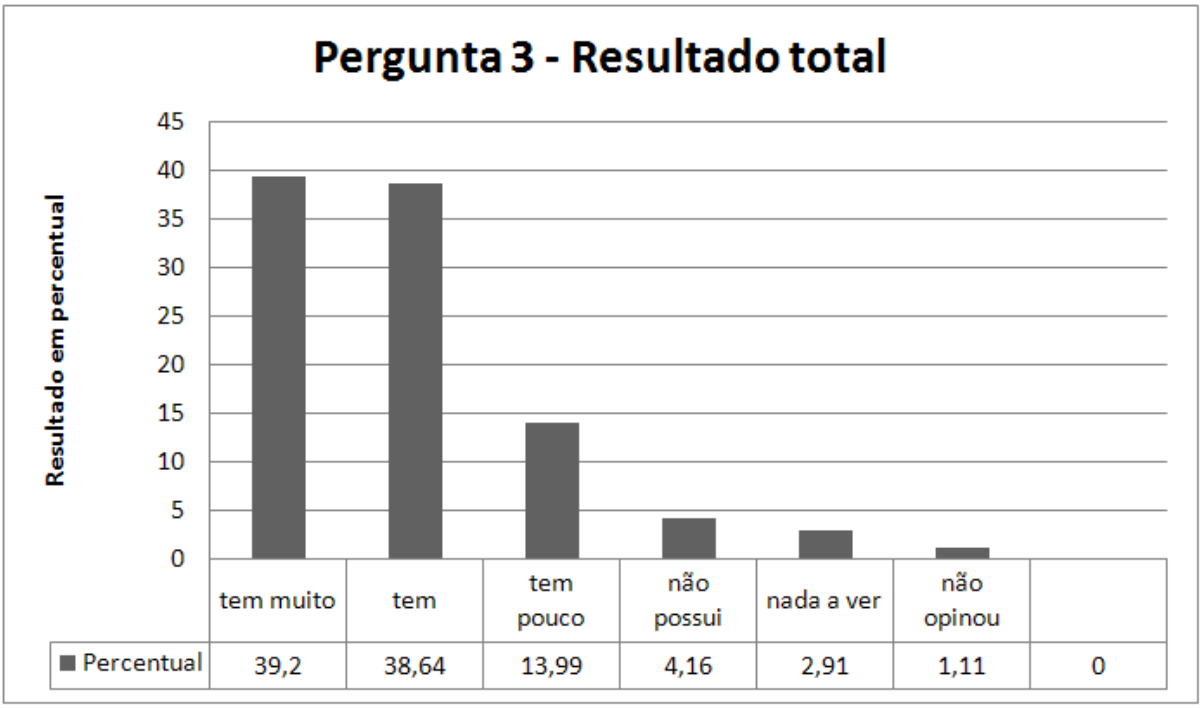


Figura 5. 10. Resultado referente às repostas da Pergunta 3 para os três assuntos usados no estudo de caso.

Fazendo uma comparação entre os resultados obtidos entre a Pergunta 1 e a Pergunta 2 considerando apenas as duas primeiras alternativas (tem muito e tem), a Pergunta 1 possui uma certa vantagem, isto é, 80.89% contra 77.84%. Isso mostra que o fato dos usuários estarem falando sobre o mesmo assunto, não significa que eles podem ter interesse por ele.

Na Figura 5.11 é apresentada a diferença entre o "uso" de conhecimento cultural na busca e o "não uso", em relação a identificação de um possível interesse do usuário pelo assunto em questão.

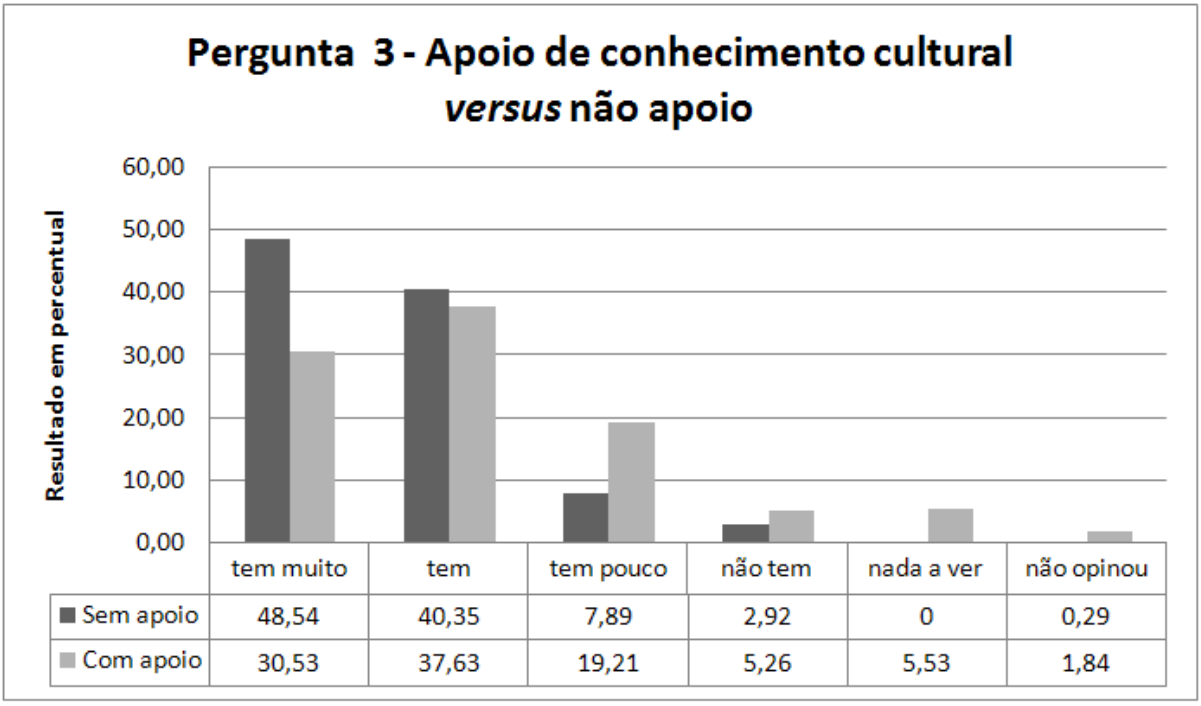


Figura 5. 11. Comparação entre o resultado do "uso" e o "não uso" de conhecimento cultural em relação às respostas da Pergunta 3.

Nota-se que o "não uso" de conhecimento cultural leva vantagem quanto a esse quesito. Identificando postagens onde seu conteúdo sugere que os usuários demonstraram mais interesse pelo assunto, isto é, 88.89% contra 68.16%. Isso considerando duas primeiras alternativas dessa pergunta (Tem muito e Tem).

Os resultado obtidos com essa pergunta, segundo a opinião dos participante do estudo de caso que se basearam apenas no teor do assunto em questão nas postagens, sugerem que buscas textuais semânticas podem considerar o contexto de uma busca, além disso, quando considera a cultura das pessoas, o resultado pode ser considerado satisfatório (Figura 5.11). Isso mostra indícios que a terceira hipótese levantada por esse estudo de caso foi alcançada como verdadeira.

5.8.4 Análise sobre a Pergunta 4

Abaixo é apresentada a Pergunta 4:

"*Se o assunto da oração (em vermelho acima) fosse de seu interesse, e se você e a pessoa que escreveu o texto fossem usuários de sites de redes sociais, você estaria disposto(a) a iniciar um contato com ele(a) ? Por quê?*

( ) Muito Disposto

( ) Disposto

( ) Pouco disposto

( ) Não disposto

( ) De modo nenhum

( ) Não tenho como opinar.

Por quê?___"

Essa pergunta é muito subjetiva e não faz parte do escopo de estudo deste trabalho. A sua elaboração se dá apenas pelo fato de identificar uma aplicabilidade para o método proposto aqui.

Com as respostas conseguidas por ela, pretende-se avaliar se o método é capaz de auxiliar Sistemas de Recomendação Social (TERVEEN, 2005) a fazer recomendação de pessoas. Quando aplicada a Pergunta 4 foi pedido aos participantes, além de estar explícito na pergunta, que eles a respondessem simulando o uso de um Sistema de Recomendação de Pessoas, onde os assuntos usados na busca representassem seus interesses.

Na Figura 5.12 são apresentados os dados referentes aos três assuntos usados na busca. Considerando as duas primeiras alternativas da Pergunta (Muito disposto e Disposto), observou-se que 63.02% das postagens recuperadas pelo método, convenceram, pelo seu conteúdo, os participantes do estudo de caso em iniciar um contato social com os usuários responsáveis por elas.

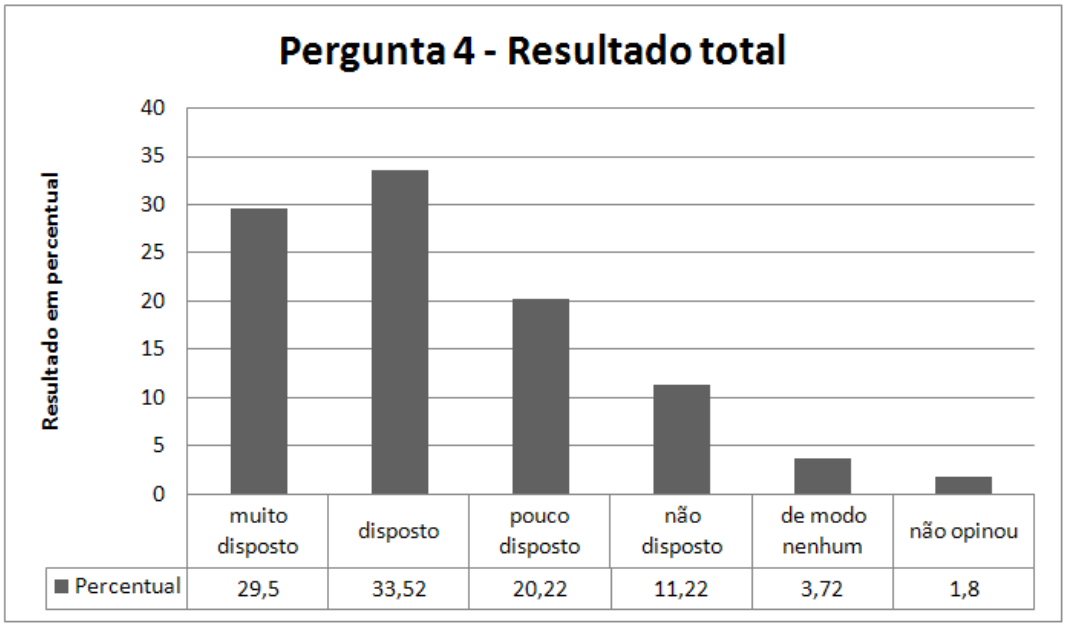


Figura 5. 12. Resultado referente às repostas da Pergunta 4 para os três assuntos usados no estudo de caso.

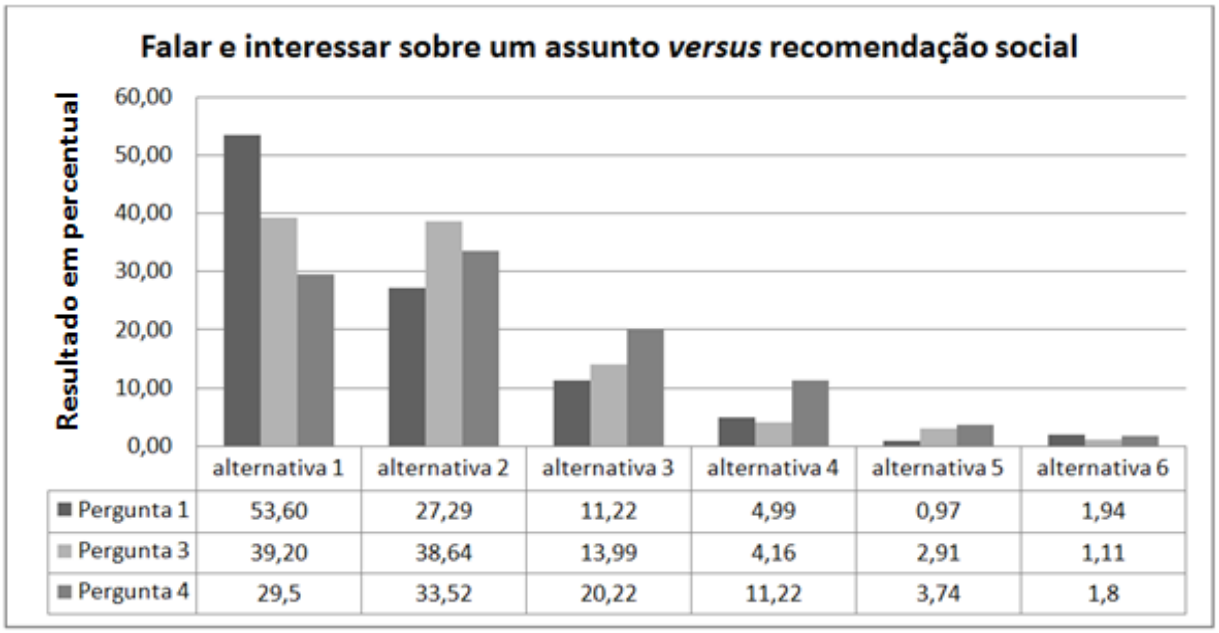
A Pergunta 4 também tinha um local ("Porque?") onde os participantes tinham a oportunidade de fazer um comentário sobre a sua escolha entre qualquer uma das alternativas. Alguns comentários negativos para essa pergunta estão na Tabela 5.6, como também o perfil e a resposta do participante.

Tabela 5. 6. Comentários de alguns participantes do estudo de caso quando respondiam a Pergunta 4.

Perfil do participante	**Resposta**	**Comentário ("Porque?")**
30-45 anos Segundo grau completo Lê uma vez por dia Não usa SNS	De modo nenhum	"*Ele é muito chato e ignorante*"
30-45 anos Segundo grau completo Lê uma vez por dia Não usa SNS	De modo nenhum	"*Ele está sendo irônico*"
30-45 anos Superior completo Lê uma vez por semana Usa SNS (Orkut)	Não disposto	"*Ele está tirando o sarro e fala muito palavrão*"
18-29 anos Superior incompleto Lê uma vez por semana Usa SNS (Orkut)	De modo nenhum	"*muito sem graça, a pessoa fala muito e é muito repetitiva.*"
18-29 anos Segundo grau completo Lê uma vez por mês Usa SNS (Orkut)	Não disposto	"*muito mal educado*"

Além desses comentários apresentados na Tabela 5.6, existem muitos outros relacionados à negação do participante, afinal 14.96% (Figura 5.12) das respostas dos participantes rejeitaram uma possível ligação social com o responsável pela postagem avaliada.

Esses dados mostram o quanto é particular a escolha de uma pessoa sobre iniciar uma ligação social com outra, pois, aconteceram casos em que os participantes consideram que a pessoa que escreveu certa postagem estava falando sobre o assunto (resposta - concordo fortemente para a Pergunta 1), e avaliando a mesma postagem disseram que não iniciaria uma ligação social com o responsável pela postagem (resposta – de modo nenhum da Pergunta 4).



	alternativa 1	alternativa 2	alternativa 3	alternativa 4	alternativa 5	alternativa 6
Pergunta 1	53,60	27,29	11,22	4,99	0,97	1,94
Pergunta 3	39,20	38,64	13,99	4,16	2,91	1,11
Pergunta 4	29,5	33,52	20,22	11,22	3,74	1,8

Figura 5. 13. Comparação entre o resultado da Pergunta 1, Pergunta 2 e Pergunta 3.

Na Figura 5.13 observa-se que os participantes do estudo de caso consideraram que 80.89% (concordo fortemente e concordo) das postagens avaliadas representam que os usuários estavam falando sobre um mesmo assunto; 77.84% (tem muito e tem) das postagens representam que os usuários tinham interesse sobre um assunto, entretanto, quando foi perguntado aos participantes se eles se interessariam em iniciar uma ligação social 60.01% (muito disposto e disposto) disseram que sim.

Esses resultados, apresentados na Figura 5.13, são um indício de que apenas identificar pessoas que estão falando sobre o mesmo assunto, ou que se interessam sobre o mesmo assunto, não significa que elas podem iniciar uma ligação social.

Esses dados sugerem que se devem considerar outros fatores quando há a intenção de recomendar pessoas uma a outra, ou seja, não apenas as comparações textuais, seja ela direta ou semântica, contudo, mesmo assim este trabalho considera que 60.01% é um número razoável quando se lida com assunto que possui subjetividade.

5.9 Considerações finais

Nesse capítulo foi descrito um estudo de caso, em que houve participação de 19 pessoas, a fim de observar o resultado do método proposto por este trabalho.

Os resultados apresentados sugerem que comparações semânticas textuais conseguem identificar pessoas que estão falando sobre o mesmo assunto em SNSs, como também, conseguem manter o contexto de uma busca.

Ao observar os resultados também foi possível perceber que quando se pretende recomendar pessoas uma as outras em SNS, deve considerar algo mais do que somente um único parâmetro, como é o caso deste trabalho com comparações semânticas textuais. Outras observações e trabalhos futuros deste trabalho são descritos no próximo capítulo.

6 CONCLUSÕES E TRABALHOS FUTUROS

6.1 Síntese dos principais resultados

Este trabalhou apresentou um método que se propõe a identificar pessoas que estão falando sobre o mesmo assunto em SNSs. A sua principal contribuição está relacionada ao uso de uma base cultural com o intuito de identificar e considerar a cultura das pessoas nas buscas feitas para identificá-las.

O uso do conhecimento cultural se tornou um fator indispensável, apesar de alguns problemas relatados, para o propósito do método, pois por meio dele é possível identificar pessoas que estão falando sobre o mesmo assunto, mesmo que de forma diferente, ao contrário dos métodos tradicionais de pesquisa, como por exemplo o do Orkut e do Facebook, que fazem comparações de palavras.

Com os resultados obtidos através do estudo de caso foi possível perceber que o uso do método é viável para o que ele se propõe a fazer, ou seja, identificar pessoas que estão falando sobre o mesmo assunto. Entretanto, é válido mencionar que quando se trata de recomendações de pessoas, percebe-se que outros fatores devem ser considerados para um melhor desempenho. Fatores esses que não foram identificados com tanta precisão por este trabalho, pois apenas percebeu-se que há outros fatores, que não somente o texto, que influenciam as pessoas a incluir ou não outras em sua rede de contatos.

Outro ponto importante a se ressaltar sobre o trabalho de pesquisa é a forma como ele foi conduzido. A abordagem de trabalho adotada mostrou-se útil, principalmente em apontar as falhas, que em alguns casos poderiam ser imperceptíveis sem os experimentos preliminares previstos.

Outro ponto que há de se destacar na abordagem de trabalho adotada é a forma evolucionária, inspirada no modelo espiral de Engenharia de Software (BOEHM, 1986), de melhoramento contínuo da proposta. Por meio desse mecanismo foi possível partir de uma proposta um tanto quanto ingênua para alcançar uma que apresentou bons resultados.

A principal dificuldade percebida durante o desenvolvimento da pesquisa foi em relação ao uso de ferramentas de apoio ao método, como por exemplo, o *parser* PALAVRAS. Apesar dessa ferramenta ser considerada o melhor analisador sintático para o português do Brasil, apresentou alguns erros, não em relação a sua análise, mas quanto a disponibilização dos resultados da análise, pois, os documentos ".XML" gerados por ele não

eram “bem formados” como era especificado em sua documentação, obrigando a realizar correções que poderiam ser evitadas caso o *parser* funcionasse adequadamente.

Outra dificuldade extremamente importante durante os trabalhos foi a respeito da mineração de dados no Orkut. Como este trabalho não tem acesso ao banco de dados desse SNS, foi necessário implementar um sistema para minerar os dados das páginas web da aplicação, perdendo um tempo valioso com o desenvolvimento de um analisador, ou melhor, um *parser* de HTML somente para minerar dados.

O grande problema é que no segundo semestre de 2009 e de 2010 o SNS Orkut mudou toda a forma de apresentação dos dados dos usuários, obrigando que o aplicativo desenvolvido para os testes fosse alterado. Problema esse, que poderia ser evitado se houvesse acesso a um banco de dados de um SNS.

6.2 Trabalhos futuros

Apesar do resultado satisfatório do estudo de caso, observou-se que algumas melhorias podem ser feitas para aperfeiçoar o método. Uma delas é melhorar a busca utilizando o conhecimento cultural na base do OMCS-Br, pois considerando apenas um conceito na busca podem ser obtidos resultados indesejáveis ao assunto em questão.

Por exemplo, na base cultural podem estar armazenados dois conhecimentos, que são distintos e, isso pode prejudicar o bom funcionamento do método. Um deles *IsA(Lula, presidente)* é tido como inerente ao assunto política, quando se pretende expandir semanticamente uma *mr defender (Lula, Sarney),* como foi o caso do estudo de caso deste trabalho, mas o segundo *IsA(lula, molusco)* não faz sentido ao assunto em questão, pois a expansão da *mr* ficaria *defender (molusco, Sarney).*

Esse tipo de problema foi percebido e deve ser considerado em trabalhos futuros, pois dependendo do assunto, ele pode trazer influências ruins ao método. Mesmo quando existe uma probabilidade muito baixa de encontrar a *mr* “defeituosa” em uma postagem, como foi o caso do estudo de caso apresentado por este trabalho.

Em relação a utilizar o método para apoiar Sistemas de Recomendação Social, deve-se evoluí-lo para considerar não somente o que um usuário escreve sobre um determinado assunto, mas também outros aspectos que consiga identificar com maior precisão o interesse do usuário, como por exemplo, verificar se as comunidades que o usuário está inscrito têm algo a ver com o assunto que conseguiu identificá-lo.

Por fim, pretende-se investigar o uso do método aqui proposto em outros contextos, como por exemplo, no meio empresarial para identificar pessoas que falam sobre o mesmo negócio, ou até mesmo possuem interesse por ele, além disso, pretende-se investigar o uso do método em contextos voltados para o Marketing, pois poderia identificar pessoas que demonstram interesse por certo tipo de produtos.

Uma outra possibilidade é utilizar o método para identificar conteúdos na web, sejam eles conteúdos para aprendizagem, vídeos, fotos, etc., pois a forma com que uma pessoa nomeia um arquivo e o salva pode ser influenciada pela sua cultura e, isso pode dificultar uma outra pessoa na localização do mesmo. A ideia é semelhante ao que já foi descrito neste trabalho, por exemplo, uma pessoa pode salvar uma figura do Lula como "presidente do Brasil" e uma outra pessoa poderia fazer buscas de figuras utilizando apenas a palavra Lula e, por isso, ter dificuldade em encontrar outras figuras do mesmo contexto com outros nomes. Ao usar a base de conhecimento de senso comum teria a chance de saber que Lula é o "presidente do Brasil" é com isso o usuário também teria acesso a essa imagem.

7 REFERÊNCIAS

ALMEIDA, M. BAX, M. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, v. 32, n. 3, 2004. ISSN 1518-8353.

ANACLETO J. C., FERREIRA. A. M., PEREIRA. E. N., SILVA. M. A. R., FABRO. J. A. Ambiente para criação de jogos de cartas educacionais contextualizados. In: *WIE - Workshop sobre Informática na Escola*, 2008.

ANDERSON, R. E. Social impacts of computing: Codes of professional ethics. *Social Science Computer Review*, v. 10, n. 4, p. 453-469, 1992.

BICK, E. The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Department of Linguistcs. University of Arhus, DK, 2000. 505p.

BIRD, C. et al. Mining email social networks. In: *MSR '06: Proceedings of the 2006 international workshop on Mining software repositories*. New York, NY, USA: ACM, 2006. p. 137-143. ISBN 1-59593-397-2.

BOEHM, B. A spiral model of software development and enhancement. *SIGSOFT Softw. Eng. Notes*, ACM, New York, NY, USA, v. 11, n. 4, p. 14-24, 1986. ISSN 0163-5948.

BOLLEGALA, D. T. MATSUO, Y. ISHIZUKA, M. Measuring the similarity between implicit semantic relations from the web. In: *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, 2009. p. 651-660. ISBN 978-1-60558-487-4.

BOYD, D.; ELLISON, N. B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, v. 13, n. 1-2, November 2007. Disponível em: <http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html>.

CARVALHO, A. F. P. *Utilização de Conhecimento de Senso Comum no Planejamento de Ações de Aprendizagem Apoiado por Computador*. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, UFSCar, 2007, 257p.

CHEN, J. et al. Make new friends, but keep the old: recommending people on social networking sites. In: CHI '09: *Proceedings of the 27th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2009. p. 201-210. ISBN 978-1-60558-246-7.

CHIN, J.P. DIEHL, V. A. NORMAN, K. L. Development of an instrument measuring user satisfaction of the human-computer interface. In: *Proceedings CHI'98*, 1998.

CRANDALL, D. et al. Feedback effects between similarity and social influence in online communities. In: *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2008. p. 160-168. ISBN 978-1-60558-193-4.

DIAS, C. Estudo de Caso: ideias importantes e referências, 2000. Disponível em: <www.geocities.com/claudiaad/case_study.pdf>, Janeiro 2009.

ESLICK, I. S. *Searching for commonsense*. Dissertação de Mestrado, Massachusetts Institute of Technology, MIT, 2006, 101p.

EVANS, C. The effectiveness of m-learning in the form of podcast revision lectures in higher education. *Computers & Education*. v. 50, p. 491-498, 2008.

FALBO, R. A. et al. Ontologias e Ambientes de Desenvolvimento de Software Semânticos. In: *JIISIC - Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento*. Madri, Espanha, 2004.

FERREIRA, A. M. *Ambiente de Jogos Educacionais de Adivinhação Baseados no Conhecimento de Senso Comum* . Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, UFSCar, 2008. 123p.

FIGUEIRA FILHO, F. M. GEUS, P. L, ALBUQUERQUE, J. P. Sistemas de recomendação e interação na Web Social. In: *Simpósio Brasileiro De Fatores Humanos Em Sistemas Computacionais*, Porto Alegre: PUCRS, 2008.

FININ, T. et al. Social networking on the semantic web. *The Learning Organization: An International Journal*, Emerald Group Publishing Limited, v. 12, n. 5, p. 418-435, May 2005. ISSN 0969-6474.

GRANADA, R. et al. Formação de Equipes Profissionais Através da Avaliação de Similaridade entre Currículos. In *Proceedings of the SBSI 2006*. Porto Alegre, Brasil: SBC, 2006.

HAMASAKI, M. et al. Community focused social network extraction. In: *ASWC*, 2006. p. 155-161.

HAVASI, C. SPEER, R. ALONSO, J. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: *RANLP'97: Proceeding of the Recent Advances in Natural Languages Processing*. Borovets, USA, 1997.

HOPE, T.; NISHIMURA, T.; TAKEDA, H. An integrated method for social network extraction. In: *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006. p. 845-846. ISBN 1-59593-323-9.

HOWARD, B. Analyzing online social networks. *Commun. ACM*, ACM, New York, NY, USA, v. 51, n. 11, p. 14-16, 2008. ISSN 0001-0782.

HUANG, J. RYU, P. CHOI, K. An Empirical Research on Extracting Relations from Wikipedia Text. In: 9th International Conference on Intelligent Data Engineering and Automated Learning, Daejeon, Berlin: Springer-Verlag, 2008, p. 241-249.

INSTITUTO PAULO MONTENEGRO. Indicador de Alfabetismo Funcional – INAF BRASIL 2009. Disponível em: <http://www.acaoeducativa.org/images/stories/pdfs/inaf2009.pdf>. Acesso em: 09/04/2010.

INTERNATIONAL SYMPOSIUM ON SOCIAL COMPUTING AND NETWORKING (SOCIALNET'09). Título… 2, 2010, Los Alamitos, CA, USA, 2009.

JIN, Y.; MATSUO, Y.; ISHIZUKA, M. Extracting social networks among various entities on the web. In: *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*. Berlin, Heidelberg: Springer-Verlag, 2007. p. 251-266. ISBN 978-3-540-72666-1.

JOSHI, D.; GATICA-PEREZ, D. Discovering groups of people in google news. In: *HCM '06: Proceedings of the 1st ACM international workshop on Human-centered multimedia*. New York, NY, USA: ACM, 2006. p. 55-64. ISBN 1-59593-500-2.

KAUTZ, H.; SELMAN, B.; SHAH, M. The hidden web. *AI Magazine*, v. 18, n. 2, p. 27-36, 1997.

LARAIA, R. B.: *Cultura: um conceito antropológico*. 19. ed. Rio de Janeiro, RJ, Brasil: Jorge Zahar Editor. 2006.

LENAT, D. B. CYC: a large-scale investment in knowledge infrastructure. *ACM*, ACM, New York, NY, USA, v. 38, n. 11, p. 33-38, 1995. ISSN 0001-0782.

LI, H. PANG, N. GUO, S. WANG, H. Research on textual emotion recognition incorporating personality factor. In: ROBIO 2007: IEEE International Conference on Robotics and Biomimetics, Sanya, 2007, p. 2222-2227.

LIU, H. SINGH, P. ConceptNet – A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, Kluwer Academic Publishers, Hingham, MA, USA, v. 22, n. 4, p. 211-226, 2004. Issn 1358-3948.

MARTINS, R. T. et al. CURUPIRA: Um parser funcional para a língua portuguesa. São Carlos: NILC-ICMC, Universidade Estadual de São Paulo, 2002. 13 p. (NILC-TR-02-26).

MATSUO, Y. et al. Spinning multiple social networks for semantic web. In: *AAAI'06: Proceedings of the 21st national conference on Artificial intelligence*. AAAI Press, 2006. p. 1381-1386. ISBN 978-1-57735-281-5.

MAZIERO, E.G. PARDO, T. A. S. NUNES, M. G. V. (2007). *Identificação automática de segmentos discursivos: o uso do parser PALAVRAS*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, n. 305. São Carlos-SP, Agosto, 25p.

MELLI, GABOR. ConceptNet Semantic Relation Ontology. Disponível em:. <http://www.gabormelli.com/RKB/ConceptNet_Semantic_Relation_Ontology>. Acesso em: 10/7/2009.

MINSKY, M: *The society of mind*. 1. ed. New York, NY, USA: Simon & Schuster , Inc., 1986. ISBN 0-671-60740-5.

MISLOVE, A. et al. Growth of the flickr social network. In: *WOSP '08: Proceedings of the first workshop on Online social networks*. New York, NY, USA: ACM, 2008. p. 25-30. ISBN 978-1-60558-182-8.

MIT. CapableOf. Disponível em: <http://pedia.media.mit.edu/CapableOf>. Acesso em: 09/7/2009.

MIT. CapableOfReceivingAction. Disponível em: <http://pedia.media.mit.edu/CapableOf>. Acesso em: 09/7/2009.

MOTOYAMA, M.; VARGHESE, G. I seek you: searching and matching individuals in social networks. In: *WIDM '09: Proceeding of the eleventh international workshop on Web information and data management*. New York, NY, USA: ACM, 2009. p. 67-75. ISBN 978-1-60558-808-7.

MUNIZ, M. C. M. *A construção de recursos linguistic-computacionais para o português do Brasil: o projeto Unitex-PB*. Dissertação de Mestrado, Programa de Pós Graduação em Ciências da Computação e Matemática Computacional, ICMC-USP, 2004, 92p.

NUNES, M. A. S. N.; CERRI, S. A.; BLANC, N. Towards user psychological profile. In: *IHC '08: Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems*, Porto Alegre, Brazil: Sociedade Brasileira de Computação, 2008. p. 196-203. ISBN 978-85-7669-203-4.

PREECE, J. ROGERS, Y. SHARP, H. *Interaction design: beyond human-computer interaction*. USA: John Wiley & Sons, Inc. 2002. 519p.

RHEINGOLD, H. *La Comunidad Virtual: Una Sociedad sin Fronteras*. Barcelona, Espanha: Gedisa Editorial (Colección Limites de La Ciência), 1994.

SABA, W. S. Language, logic and ontology: Uncovering the structure of commonsense knowledge. *International Journal of Human Computer Studies*, Duluth, v. 65, n. 7, p. 610-623, 2007.

SILVA, M. A. R. SILVA, J. C. A. Promoting Collaboration through a Culturally Contextualized Narrative Game. In *Proceedings of the 11th International Conference Enterprise Information Systems (ICEIS)*. Milan, Italy: Springer, Heidelberg, 2009. p. 870-881.

SILVA, M. A. R. *Uso de Senso Comum no Apoio a Jogos Narrativos para Crianças em Idade Escolar*. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, UFSCar, 2009, 122p.

SINGH, P. BARRY, B. LIU, H. Teaching Machines about Everyday Life. *BT Technology Journal*, Kluwer Academic Publishers, Hingham, MA, USA, v. 22, n. 4, p. 227-240, 2004. ISSN 1358-3948.

SINGH, P. et al. Open Mind Common Sense: Knowledge Acquisition from the General Public. In: *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*. London, UK: Springer-Verlag, 2002. p. 1223-1237. ISBN 3-540-00106-9.

SINGH, P. The public acquisition of commonsense knowledge. In: *AAAI Spring symposium on acquiring (and using) linguistic (and world) knowledge for information access*, AAAI Press, Palo Alto, Canada, Palo Alto, Canada, 2002.

SYDDANSK UNIVERSITET. Portuguese VISL symbol set. Disponível em: <http://visl.sdu.dk/visl/pt/info/symbolset-manual.html>. Acesso em: 09/7/2009.

TANG, J. et al. Extraction and mining of an academic social network. In: *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008. p. 1193-1194. ISBN 978-1-60558-085-2.

TENENBAUM, A.: *Estruturas de dados usando C*. 1. ed. São Paulo, SP, Brasil: Makron Books Editora, 1995. ISBN 8534603480.

TERVEEN, L. MCDONALD, D. W. Social matching: A framework and research agenda. *ACM Trans. Comput.-Hum. Interact*, ACM, New York, NY, USA, v. 12, n. 3, p. 401-434, 2005. ISSN 1073-0516.

TSUTSUMI, V. P. *Uso de senso comum na detecção das diferenças culturais no contexto do projeto Open Mind Common Sense*. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, UFSCar, 2006, 118p.

TYLER, J. R.; WILKINSON, D. M.; HUBERMAN, B. A. *Email as Spectroscopy: Automated Discovery of Community Structure within Organizations*. 1501 Page Mill Road, Palo Alto CA, 94304. Disponível em: <http://www.hpl.hp.com/research/idl/papers/email/email.pdf>.

WAINER, J. Métodos de pesquisa quantitativa e qualitativa para a ciência computação. *In: KOWALTOWSKI, Tomasz; BREITMAN, Karin. (Org.). Atualização em informática 2007*. Rio de Janeiro, Brasil: Sociedade Brasileira de Computação e Editora PUC-Rio, 2007. p. 221-262.

WAZLAWICK, R. S. *Metodologia de Pesquisa para Ciência da Computação*. 1. ed. Rio de Janeiro, BR: Campus, 2009. 184 p. ISBN 8535235221.

WEAVER, A. C.; MORRISON, B. B. Social networking. *Computer*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 41, n. 2, p. 97 - 100, 2008. ISSN 0018-9162.

WORKSHOP SOBRE ASPECTOS DA INTERAÇÃO HUMANO-COMPUTADOR PARA A WEB SOCIAL (WAIHCWS). Título… 2, 2010, Belo Horizonte, MG, Brasil.

YIN, R. K. *Case Study Research: Design and Methods, Third Edition, Applied Social Research Methods Series*, v. 5. 3rd. ed. California, USA: Sage Publications, Inc, 2002. 200 p. ISBN 0761925538.

APÊNDICE A – UM ESTUDO SOBRE AS RELAÇÕES DE MINSKY[20]

A.I Considerações Iniciais

Neste apêndice será apresentado um estudo detalhado sobre cada uma das vinte relações de Minsky que auxiliam no armazenamento de conhecimento cultural colhido pelo Projeto OMCS-Br. Aqui não serão apresentadas as vinte relações negativas, pois cada uma delas diferem de sua relação correspondente apenas pela negação.

A.II Relações de Minsky

As vinte relações de Minsky que fazem parte do Projeto OMCS-Br são apresentadas na Tabela I.I. Elas são divididas em oito classes que visam mapear toda e qualquer forma de conhecimento humano (TSUTSUMI, 2006).

Tabela I.I Relações semânticas que compõem o projeto OMCS-Br (LIU, 2004).

Classe	Relação	Exemplo
K-Lines	ConceptuallyRelatedTo	(ConceptuallyRelatedTo 'bad breath' 'mint' 'f=4;i=0)
	ThematicKLine	(ThematicKLine 'weding dress' 'veil' 'f=9;i=0)
	SuperThematicKLine	(SuperThematicKLine 'western civilisation' 'civilisation' 'f=0;i=12)
Things	IsA	(IsA 'horse' 'animal' 'f=17;i=3)
	PropertyOf	(PropertyOf 'fire' 'dangerous' 'f=17;i=1)
	PartOf	(PartOf 'butterfly' 'wing' 'f=3;i=0)
	MadeOf	(MadeOf 'bacon' 'pig' 'f=3;i=0)
	DefinedAs	(DefinedAs 'meat' 'flesh of animal' 'f=2;i=1)
Agents	CapableOf	(CapableOf 'dentist' 'pull tooth' 'f=4;i=0)
Events	PrerequisiteEventOf	(PrerequisiteEventOf 'read letter' 'open envelope' 'f=2;i=0)
	FirstSubeventOf	(FirstSubEventOf 'start fire' 'light match' 'f=2;i=3)
	SubEventOf	(SubEventOf 'play sport' 'score goal' 'f=2;i=0)
	LastSubeventOf	(LastSubEventOf 'attend classical concert' 'aplaud' 'f=2;i=1)
Spatial	LocationOf	(LocationOf 'army' 'in war' 'f=3;i=0)
Causal	EffectOf	(EffectOf 'view video' 'entertainment' 'f=2;i=0)
	DesirousEffectOf	(DesirousEffectOf 'sweat' 'take a shower' 'f=3;i=1)
Functional	UsedFor	(UsedFor 'fire place' 'burn' 'f=1;i=2)
	CapableOfReceivingAction	(CapableOfReceivingAction 'drink' 'serve' 'f=0;i=14)
Affective	MotivationOf	(MotivationOf 'play game' 'compete' 'f=3;i=0)
	DesireOf	(DesireOf 'person' 'not be depressed' 'f=2;i=0)

[20] Este trabalho foi realizado por Gilberto Astolfi, Johana Maria Rosas Villena e David Buzatto, alunos de pós graduação em UFSCar (Ciência da Computação da Universidade Federal de São Carlos).

A.III Classe K-lines

A classe das *K-lines* mapeia a relação entre dois conceitos que não está precisamente explicito na relação (LIU, 2004). Essa classe é composta pelas relações *ConceptuallyRelatedTo, ThematicKLine* e *SuperThematicKLine*.

ConceptuallyRelatedTo – É um tipo de relação que diz que existe uma relação entre os dois conceitos, mas não é possível determinar qual é (HAVASI, 1997), ou seja, eles são relacionados, mas por um caminho desconhecido. Exemplo: *ConceptuallyRelatedTo (Mau hálito, hortelã).* Apenas quem leu a frase de onde eles foram extraídos sabe dizer o tipo de relacionamento.

ThematicKLine – Define um relacionamento entre coisas sobre o mesmo tema, ou seja, alguma coisa que lembra outra (LIU, 2004). Exemplo: *ThematicKline (Champanhe, gelo).*

SuperThematicKLine – Unifica o tema com suas variações (LIU, 2004). Exemplo: "*Lançamento*" é um super tema para "*Lançamento de filme*" e "*Lançamento de carro*", *SuperThematicKline (Lançamento, filme)* e *SuperThematicKline (Lançamento, carro).*

A.IV Classe Things

Essa classe é composta pelas relações *IsA, PropertyOf, PartOf, MadeOf,* e *DefinedAs*.

IsA – Considerada uma relação fraca (LIU, 2004). Esta relação tem como objetivo especializar certa coisa com sentido hierárquico (LIU, 2004). Em outras palavras, "*X é um Y*" geralmente significa que o conceito *X* é uma especialização do conceito *Y,* e *Y* é um conceito generalizado do conceito *X*. "*Maçã é uma fruta*", neste exemplo "*Maçã*" é uma especificação de uma "*fruta*" e "*fruta*" é uma generalização de "*Maçã*". Exemplo: *IsA(maça, fruta).*

PropertyOf – Essa relação mapeia a possibilidade de atribuir propriedade a certa coisa. Ela é considerada uma relação com as seguintes características (MELLI, 2009): (1) direta: *Y* é uma propriedade de *X*. Exemplo: "*Salgada é uma propriedade de água*"; (2) irreflexiva: *Y* não é uma propriedade *Y,* ou seja, uma coisa não é propriedade dela mesma. Exemplo: "*Água não é uma propriedade de água*"; (3) anti-simétrica: *Y* é uma propriedade de *X* então *X* não é uma propriedade *Y*. Exemplo: "Salgada *é uma propriedade de água então Água não é uma propriedade de salgada*".

PartOf - Essa relação possibilita que uma coisa possa ser determinada como parte de outra. Ela é definida como uma relação (MELLI, 2009): (1) direta: *Y* é uma parte de X. Exemplo: "*Roda é uma parte de um carro*"; (2) irreflexiva: *Y* não é parte de *Y*, ou seja, uma coisa não é parte dela mesma. Exemplo: "*Roda não é parte de uma roda*"; (3) anti-simétrica: *Y* é uma parte de *X* então *X* não é uma parte de *Y*. Exemplo: "*Roda é uma parte de um carro então Carro não é uma parte de uma roda*"; (4) transitiva: *Y* é parte de *X* e *X* é parte de *Z* então *Y* é parte de *Z*. Exemplo: "*Calota é parte de uma roda e uma roda é parte de um carro então uma Calota é parte de um carro*".

MadeOf – É uma relação implícita de todo (*X, Y*), combinação a qual é possível quando *X* é um subtipo de *Y*, ou seja, *X* é o produto e *Y* a substancia (SABA, 2007). Exemplo: "*bacon (X) é feito de porco (Y)*", ou seja, *X* é obtido através de um processamento de *Y*.

DefinedAs – é um tipo de relação que faz uso de sinônimos para representar o significado de algo. Ela é representada da seguinte forma: *DefinedAs(X, Y)*, onde *X* é uma conceito que tem a mesma natureza que o conceito *Y*. O que se quer dizer é que *X* possui todas as características de *Y* e vice-versa (LIU, 2004). Por exemplo, *DefinedAs(Linda, Maravilhosa)*.

A.V Classe Agents

Essa classe é composta pela única relação *CapableOf*.

CapableOf – essa relação mapeia as habilidades e/ou capacidades de algo ou alguém (MIT, 2005). Por exemplo: "*Um herói é capaz de lutar*" – *CapableOf(herói, lutar)*.

A.VI Classe Events

Essa classe é composta pelas relações *PrerequisiteEventOf, FirstSubeventOf, SubEventOf* e *LastSubeventOf*.

PrerequisiteEventOf – essa relação especifica o evento obrigatório *X* que deve acorrer para que o evento *Y* seja satisfeito (MELLI, 2009). Por exemplo: *PrerequisitEventOf (passar vestibular, cursar graduação)*.

FirstSubeventOf - essa relação especifica o primeiro evento *X*, em um conjunto de eventos, relacionado a um evento final *Y*, ou seja, um evento inicial que chega a um evento final, mas entre os dois eventos podem ocorrer outros eventos (MELLI, 2009). Por exemplo: *FirstSubeventOf (iniciar jogo, ganhar)*.

SubEventOf - essa relação especifica o evento *X* que acontece depois de certo evento *Y* (MELLI, 2009). Exemplo: *SubeventOf (acordar, tomar café).*

LastSubeventOf - essa relação especifica o penúltimo evento *X*, em um conjunto de eventos, relacionado a um evento final *Y* (MELLI, 2009). Exemplo: *LastSubEventOf (quebrar objeto, ser punido).*

A.VII Classe Spatial

Essa classe é composta pela única relação *LocationOf*.

LocationOf – essa relação representa a localização espacial de algo (LIU, 2004) e tem como característica ser transitiva, isto é, *Y* é localizado em *X* e *X* é localizado em *Z* então *Y* é localizado em *Z*. Exemplo: "*Sal é localizado na água e água é localizada na comida então na comida é localizado sal*" – *LocationOf(sal, água); LocationOf(água, comida)* e *LocationOf(sal, comida).*

A.VIII Classe Causal

Essa classe é composta pelas duas relação *EffectOf* e *DesirousEffectOf* .

EffectOf – é uma relação que tipicamente modifica uma ação e é composta de pares de verbos ou verbos mais substantivos. É consequência de uma ação ou um evento (ESLICK, 2006). Ela possui a característica de transitividade, isto é, *Y* é efeito de *X* e *X* é efeito de *Z* então *Y* é efeito de *Z*. Exemplo: "*Calafrio é efeito da febre e febre é efeito da gripe então calafrio é efeito da gripe*".

DesirousEffectOf – essa relação mapeia um possível efeito ou conseqüência do desejo por alguma coisa (ESLICK, 2006; LI, 2007). Ela também possui a característica de transitividade, isto é *Y* é um efeito desejável de *X* e *X* é um efeito desejável de *Z* então *Y* é um efeito desejável de *Z*. Exemplo: "*Ficar rico é um efeito desejável de trabalhar e trabalhar é um efeito desejável de ganhar dinheiro então ficar rico é um efeito desejável de ganhar dinheiro*".

A.IX Classe Functional

Essa classe é composta pelas relações *UsedFor* e *CapableOfReceivingAction*.

UsedFor – é uma relação que especifica a função de uma entidade. *X* pode ser usado para *Y*, ou *X* pode ser usado em *Y* (HUANG, 2008). Exemplo: *UsedFor(faca, cortar alimento).*

CapableOfReceivingAction – essa relação algo ou alguém que pode receber uma ação (MIT, 2005b). Exemplo: *CapableOfReceivingAction(bola, chutar).*

A.X Classe Affective

A classe *Affective* é composta pelas relações *MotivationOf* e *DesireOf*

MotivationOf – essa relação representa o efeito *Y* de algo *X*, como também a motivação de algo (LIU, 2004). Por exemplo: *MotivationOf(jogar, competir).*

DesireOf - essa relação especifica o desejo Y de algo X (ESLICK, 2006). Exemplo: *DesireOf(cachorro, comida).*

A.XI Considerações Finais

Neste apêndice foi apresentado um estudo sobre as relações semânticas que compõe o Projeto OMCS-Br. Esse estudo serviu como base para a tomada de decisão e adoção de algumas abordagens durante o desenvolvimento do método proposto por este trabalho de pesquisa.

APÊNDICE B – QUESTIONÁRIO PARA COLETA DE PERFIL

Por favor, marque com um "X" a alternativa que você considera apropriada para seu perfil.

1) Qual a faixa de idade que você se encaixa?
   ( ) Entre 13 e 17 anos
   ( ) Entre 18 e 29 anos
   ( ) Entre 30 e 45 anos
   ( ) Entre 45 e 64 anos
   ( ) Acima de 65 anos

2) Qual o seu grau de escolaridade?
   ( ) Primeiro grau incompleto
   ( ) Segundo grau completo
   ( ) Segundo grau incompleto
   ( ) Segundo grau completo
   ( ) Superior incompleto
   ( ) Superior completo
   ( ) Pós graduado (Latu Senso)
   ( ) Mestrado
   ( ) Doutorado

3) Você gosta de ler livros, revistas, jornais, ou outros?
   ( ) Sim
   ( ) Não
   Caso queira especificar que tipo de leitura gosta, escreva aqui:

4) Qual a sua frequência de leitura?
   ( ) Uma vez por dia
   ( ) Uma vez por semana
   ( ) Uma vez ao mês
   ( ) Outros – Especifique aqui:

5) Você tem conta em sites de redes sociais, como por exemplo, Orkut, Facebook, Twitter, etc?
   ( ) Sim – Especifique:
   ( ) Não

Por favor, agora responda o segundo questionário.

APÊNDICE C – QUESTIONÁRIO PARA AVALIAR O MÉTODO PROPOSTO

Por favor, para cada texto à esquerda na tabela responda as quatro questões.

Oração usada como assunto: "Aqui vai a oração usada como assunto"				
Textos ➡	Você concorda que a pessoa que escreveu o texto à esquerda está falando sobre o mesmo assunto da oração (em vermelho acima)?	Considerando somente o trecho em destaque no texto à esquerda, você acha que ele possui o mesmo significado que a oração (em vermelho acima) usada como assunto?	Você acha que a pessoa que escreveu o texto à esquerda tem interesse sobre o assunto da oração (oração em vermelho acima)?	Se o assunto da oração (em vermelho acima) fosse de seu interesse, e se você e a pessoa que escreveu o texto fossem usuários de sites de redes sociais, você estaria disposto(a) a iniciar um contato com ele(a)? Por quê?
Aqui vai a postagens recuperada do Orkut pelo método.	Concordo fortemente ( )	Possui muito ( )	Tem muito ( )	Muito disposto ( )
	Concordo ( )	Possui ( )	Tem ( )	Disposto ( )
	Concordo pouco ( )	Possui pouco ( )	Tem pouco ( )	Pouco disposto ( )
	Discordo ( )	Não possui ( )	Não tem ( )	Não disposto ( )
	Discordo fortemente ( )	Não tem nada a ver ( )	Não tem nada a ver ( )	De modo nenhum ( )
	Não posso opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( )	Não tenho como opinar ( ) Por quê? _____

Muito obrigado pela sua colaboração.