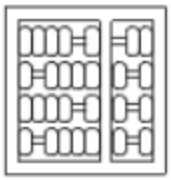

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Identificando Semântica em Redes Sociais Inclusivas Online: Um Estudo sobre Ferramentas e Técnicas

Júlio Cesar dos Reis
Rodrigo Bonacin
Maria Cecília Calani Baranauskas

Technical Report - IC-10-28 - Relatório Técnico

September - 2010 - Setembro

Identificando Semântica em Redes Sociais Inclusivas *Online*: Um Estudo sobre Ferramentas e Técnicas

Júlio Cesar dos Reis[1 2], Rodrigo Bonacin[2], M. Cecília C. Baranauskas[1]

[1]Departamento de Sistemas de Informação, Instituto de Computação
Universidade Estadual de Campinas – UNICAMP
13083-970, Campinas, SP, Brasil

[2] CTI Renato Archer - Rodovia Dom Pedro I, km 143,6, 13069-901, Campinas, SP, Brasil

{julio.reis e rodrigo.bonacin}@cti.gov.br, cecilia@ic.unicamp.br

Abstract

Access to knowledge is a basic condition for living in the digital age and Social Network Services are a reality nowadays. Search mechanisms are increasingly essential for interaction and information retrieval in such systems. Appropriate representation of the meaning that people use in SNS can be a determining factor for the development of more adequate search engines. The identification of concepts and semantic relationship that come out from the network data are even more relevant for Inclusive Social Network Services (ISN), which presuppose respect for the diversity of users, including those in the process of digital literacy. This work studies tools and techniques for the identification of concepts and semantic relationships in ISN, aiming at designing a strategy to assist the building of ontologies that model the shared semantics in the social network, toward more adequate search mechanism for ISN. The extraction process points out results which demonstrate the importance of applying appropriated methods to the considered context.

Keywords: *Inclusive Social Network Services; Semantic Search; Exploitation of Social Knowledge; Text Mining; Ontology Engineering.*

Resumo

O acesso ao conhecimento é condição básica para a vida na era digital e Rede Sociais *Online* (RSO) são uma realidade nos dias de hoje. Mecanismos de busca são cada vez mais essenciais para a interação e recuperação da informação em tais sistemas. Representações adequadas dos significados que as pessoas usam na RSO podem ser um fator determinante para o desenvolvimento de mecanismos de busca mais adequados. A identificação de conceitos e relações semânticas que advém dos dados da RSO é ainda mais relevante para Redes Sociais Inclusivas

Online (RSI), que pressupõem o respeito à diversidade de usuários, incluindo aqueles em processo de alfabetização digital. Neste contexto, este trabalho estuda ferramentas e técnicas para a identificação de termos e relações semânticas em RSI visando a concepção de uma estratégia para auxiliar na construção de ontologias que modelem a semântica compartilhada na rede social, rumo a um mecanismo de busca mais adequado a RSI. O processo de extração aponta resultados que demonstram a importância da aplicação de métodos apropriados ao contexto considerado.

Palavras-Chave: Redes Sociais Inclusivas *Online;* Busca Semântica; Conhecimento Social; Engenharia de Ontologias; Mineração de Texto.

1 Introdução

Denominamos Redes Sociais Inclusivas *Online* (RSI) as redes sociais mediadas por sistemas computacionais, nas quais cada pessoa pode integrar grupos e interagir para produzir elementos que podem ser compartilhados [1]. Estas diferem de outras Redes Sociais *Online* (RSO) pela atenção especial aos aspectos universais (no sentido de serem para todos em sua maior extensão possível), onde cada funcionalidade do sistema é desenhada considerando a diversidade e as diferenças de competência das pessoas, por exemplo, no seu letramento.

Mecanismos de busca em RSI também devem considerar esse contexto onde não se pode pressupor um usuário familiarizado com os procedimentos e, em muitos casos, com os algoritmos de busca na *Web*. Além disso, tais mecanismos deveriam considerar os significados compartilhados pelos indivíduos na rede social, podendo assim propiciar buscas mais adequadas. O conhecimento do domínio, no qual mecanismos semânticos usualmente estão fundamentados, deveria ser baseado em atividades da rede social; isto inclui dar ênfase à linguagem local e cotidiana das pessoas que utilizam a rede e se organizam em comunidades. Para isso são necessárias ferramentas e técnicas que permitam efetuar extração e mineração dos dados do sistema, de modo a descobrir e modelar a semântica compartilhada pelas pessoas na rede social [2].

Conforme vislumbrado por Berners-Lee *et al.* [3] a *Web* tende ao entendimento e uso das informações disponíveis sob o ponto de vista semântico. Na *Web* Semântica ter-se-ia conteúdo usável por máquinas, não apenas para o propósito de apresentação, mas para automação, integração e reuso entre aplicações [4]. Embora uma *Web* Semântica ainda não esteja inteiramente à disposição de seus usuários, desenvolver aplicações e serviços *Web* fundamentados em estruturas computacionais semânticas que representem o conhecimento tem sido uma meta.

Funcionalidades em aplicações da *Web* Social [5], como integração semântica entre *Wikis* e busca em sistemas de Redes Sociais tendem, cada vez mais, a utilizar destes recursos para oferecer serviços melhores. A modelagem do conhecimento para estas aplicações semânticas pode ser feita através das ontologias computacionais. Gruber [6] define

ontologia, no contexto da Ciência da Computação, como uma especificação formal de um conjunto de conceitos e suas relações, que fornece descrições sobre conhecimento.

A construção de ontologias usualmente é uma tarefa difícil e demorada que necessita combinar o conhecimento de especialistas na área, com a habilidade e experiência de engenheiros de ontologia em um esforço único. Essa dificuldade ainda é agravada pela crescente quantidade de dados disponíveis, junto à diversidade e complexidade dos assuntos. Logo, são necessárias soluções computacionais (semi)automáticas que auxiliem na construção das ontologias como um meio de representar a semântica.

A identificação de conceitos e a descoberta de relações semânticas entre conceitos é um problema abordado na área de aprendizagem de ontologias (*Ontology Learning*). A literatura relata tentativas de desenvolver aprendizagem de ontologias a partir de textos *e.g.* [7] e também a partir de marcações semânticas (*tags*) *e.g.* [8]. Todavia, as soluções ainda são intrinsecamente relacionadas e dependentes de contextos e domínios bem definidos, em idiomas específicos. Pelo fato de considerar pessoas poucos experientes com artefatos digitais, em uma RSI não se pode esperar que usuários incluam marcações semânticas em um primeiro momento, tal como o uso intensivo de termos cultos definidos em vocabulários formais. Portanto isto impõe dificuldades adicionais à mineração dos textos e a construção de ontologias neste contexto.

Neste contexto, este trabalho analisa a aplicação de possíveis ferramentas e técnicas para a identificação de termos relevantes e relações semânticas em dados de uma RSI. Este artigo apresenta um experimento elaborado com dados reais da RSI '*VilanaRede*'[1]. A escolha e análise das ferramentas e técnicas para o experimento são baseadas nas características dos dados presentes na rede, considerando que o idioma é o português e há uma grande diversidade de assuntos no conteúdo da rede. Devido à dificuldade de encontrar ferramentas de análise lingüística neste idioma, a maioria das ferramentas e técnicas utilizadas neste trabalho é de base estatística, que não necessita de anotações complexas no texto. Ferramentas especialmente projetadas ao idioma português com abordagens híbridas também são averiguadas.

Utilizando estas ferramentas e técnicas, o objetivo deste artigo é buscar uma melhor estratégia para auxiliar na construção de ontologias que representem significados utilizados e compartilhados pelos usuários da rede social, gerando a possibilidade de explorar o conhecimento social. Não se pretende verificar qual a melhor ferramenta isoladamente, nem propor um método para geração automatica de ontologias, contudo explorar como estas ferramentas podem ser utilizadas de maneira conjunta em um processo para se auxiliar a

[1] www.vilanarede.org.br

modelagem de ontologias baseadas no conhecimento presente na RSI. A identificação de termos e relações semânticas nos dados da rede propiciará meios para a construção destas ontologias. Estas serão úteis na proposta de mecanismos de busca mais adequados aos sistemas de RSI, uma vez que tais mecanismos deveriam levar em consideração os significados criados, compartilhados e usados pelas pessoas através do sistema [2] - *i.e.* os significados que as pessoas trouxeram para a rede, e também os que foram tecidos com o uso do sistema ao longo do tempo (através de interação).

Este relatório técnico está organizado da seguinte forma: A seção 2 apresenta a motivação para este estudo e os trabalhos relacionados visando contextualizar a pesquisa; na seção 3 são apresentadas as ferramentas e técnicas utilizadas no experimento, junto à justificativa da escolha destas; em seguida na seção 4 é apresentada a configuração do experimento, junto aos resultados e uma discussão; a seção 5 conclui o trabalho e aponta os trabalhos futuros.

2 Identificando e Representando os Significados Compartilhados na RSI

Atualmente, devido à imensa quantidade de informações gerada pelos usuários na *Web* têm-se necessidades e também dificuldades crescentes em gerenciar e recuperar a informação neste ambiente. Isto é um reflexo da grande diversidade cultural do mundo contemporâneo que têm tornado a informação ainda mais complexa. Estes fatores são ainda mais aparentes em aplicações de RSO, que formam comunidades com interesses comuns e há uma intensa interação social entre pessoas.

Neste cenário, soluções com foco em uma *Web* mais centrada nos aspectos humanos, observando e valorizando a diversidade, poderiam representar soluções mais adequadas a busca de informação. Em contextos de Brasil e outros países em desenvolvimento, isso se torna ainda mais necessário, devido à urgência de soluções que considerem diferentes níveis de letramento digital e que facilitem o acesso ao conhecimento.

Os aspectos da linguagem dos usuários e como estes fazem sentido das coisas no mundo deveriam ter maior atenção no desenvolvimento de soluções de *software*; isto compreende soluções que atendam a linguagem coloquial, ao regionalismo e outras riquezas da linguagem humana. O acesso à informação pode ser facilitado quando levamos em consideração estes aspectos, uma vez que os usuários, principalmente aqueles com baixo letramento, utilizam termos de sua vida cotidiana que fazem sentido a eles, mas nem sempre são parte da língua culta ou formal [2].

O conceito de RSI [1] está alinhado ao desenvolvimento de sistemas que façam sentido a uma comunidade de pessoas, objetivando incluir todos na constituição de uma cultura digital. Nesta direção, investigações sobre mecanismos de busca mais adequados a sistemas de RSI tendem a levar o aspecto da linguagem do usuário e seus significados em consideração [2], podendo gerar oportunidades e facilidades para o acesso à informação. Para tal solução, precisamos de artefatos computacionais que tenham o poder de representar a semântica da linguagem dos indivíduos da rede social, com o intuito de gerar resultados de

busca que façam sentido ao usuário que faz a busca. As ontologias *Web* podem contribuir nesta tarefa.

Para desenvolver um mecanismo de busca mais adequado ao contexto da linguagem usada pelos usuários da RSI, os dados da modelagem semântica utilizados pela busca deveriam advir da própria rede social, objetivando maior fidelidade aos termos e significados utilizados pelas pessoas da rede [2]. A extração de termos e possíveis conceitos junto a descoberta de relações semânticas são passos fundamentais e necessários para a representação semântica, construindo ontologias com os significados compartilhados na rede. De modo geral, a construção de ontologias passa por diversas etapas como: extrair conceitos, extrair relações (taxonômicas e não-taxonômicas) e popular a ontologia com instâncias.

Segundo Šimko & Bieliková [9] a identificação de conceitos e seus relacionamentos é crucial para a criação de ontologias, mas infelizmente, a identificação manual é uma tarefa tediosa mesmo para pequenos domínios. Para domínios dispersos e com grande volume de dados esta tarefa se torna ainda mais custosa e demorada, o que inviabiliza a sua execução manual. Se existem dezenas de conceitos identificados, as relações podem ser centenas ou milhares.

Os métodos de extração de conceitos podem variar, sendo os conceitos extraídos a partir dos documentos como os termos mais relevantes disponíveis no conteúdo. Além dos termos, as abordagens relacionadas às relações criadas utilizam principalmente técnicas de processamento de linguagem natural. A extração de relações semânticas usualmente também faz uso da análise de correlação de termos para identificar agrupamentos de termos que estejam conceitualmente próximos. De acordo com Šimko & Bieliková [9] nas abordagens existentes, a descoberta de relacionamentos é induzida com base principalmente na análise lingüística precisa, invocando anotação de texto precedente, sendo esta sua principal desvantagem. A maioria das abordagens depende de anotações lexicais ou sintáticas, necessitam de poderosos mecanismos de anotação de texto, são dependentes de ontologias de domínio existentes, ou de recursos semânticos externos (*e.g. WordNet*[2]).

Devido à complexidade no desenvolvimento das ontologias, a investigação sobre etapas de extração de termos e relações semânticas pode levar a melhores resultados na sua construção. Existem diversas propostas na literatura relacionadas à construção automática de ontologias (aprendizagem de ontologias) a partir de textos *e.g.* [7], contudo muitas soluções ainda são altamente dependentes de domínios restritos e de idiomas. A maioria das

[2] http://wordnet.princeton.edu/

propostas é baseada em *corpus* de domínio, que é um conjunto de textos sobre determinada área do conhecimento (*e.g.* Medicina).

Neste trabalho os textos são conteúdo da RSI '*Vila naRede'*. Os usuários desta RSI interagem com o sistema e entre si principalmente criando anúncios de produtos, serviços e idéias. Eles têm a possibilidade de se comunicarem através de ferramentas especificas e através de comentários nos anúncios. Os usuários desta RSI são majoritamente pessoas brasileiras, falantes da língua portuguesa, e em fase de letramento digital. O principal desafio no contexto do '*VilanaRede'* é lidar com um domínio não fechado no conteúdo dos anúncios. O conteúdo criado a partir deste sistema forma um *corpus* de referência, que é constituído por conteúdos de diversas áreas e assuntos, tornando-se independente de um domínio específico. Assim o desafio está em efetuar o processamento de texto em um domínio não fechado, com a dificuldade de lidar com contextos diversos.

Aprendizado de ontologias a partir de sistemas sociais tem sido explorado em alguns trabalhos na literatura (*e.g 8*, 11, 12, 24, 25). Contudo, um estudo no contexto de RSI ainda não foi explorado na literatura. Uma vez que o objetivo de uma RSI é ser para todos, o conteúdo presente nestes sistemas (*i.e.* os anúncios) tende a ser heterogêneo com relação a: assunto dos conteúdos, formas de expressão, linguagem dos usuários, idade, limitações físicas e sensoriais, identidade, entre outros fatores. Estes anúncios são diversificados quanto ao conteúdo, como: a venda de produtos artesanais variados, comida, produtos de informática, serviço de advocacia, eventos incluindo debate sobre educação e festas juninas; assim como idéias que incluem receitas, conscientização ambiental, dicas de saúde, entre outros.

Tal fato tende dificultar o êxito dos métodos clássicos de extração de ontologias dependentes de línguas específicas e contextos bem definidos e também de redes sociais tradicionais. A diversidade de prospectivos usuários de uma RSI pode ser demonstrada segundo estatísticas de órgãos de pesquisa como o IBGE[3]. Trabalhos como [1, 2] descrevem algumas características desta diversidade.

O estudo sobre ferramentas e técnicas para a extração de semântica (termos e relações), a partir destes dados da RSI poderão gerar ontologias mais representativas, que serão utilizadas por um mecanismo semântico de busca na rede. Até onde é de nosso conhecimento não há na literatura estudos que compreendam esta análise, sugerindo a melhor abordagem em termos de ferramentas ou técnicas de extração para este contexto. O objetivo deste trabalho é analisar algumas ferramentas e técnicas de extração de termos e relações semânticas, presentes na literatura, para criar a ontologia a partir dos dados reais da

[3] http://www.ibge.gov.br/home/

rede social. A próxima subseção sintetiza da literatura algumas técnicas e experimentos sobre extração de semântica.

2.1 Experimentos Sobre Extração de Semântica

Segundo Alema-Meza *et al.* [10] diversos estudos têm explorado a idéia de redes sociais para desenvolver técnicas de extração de semântica. Pesquisas mais significativa são principalmente de [8, 11, 12]. O trabalho de Mika [8] utiliza a idéia de redes sociais para a construção de ontologias.

Mika desenvolveu um sistema para a extração, agregação e visualização de redes sociais *online* para uma comunidade na *Web* Semântica. Mika [8] discute as maneiras como uma comunidade se desenvolve e os seus compromissos de mudança: *e.g.* porque seus membros saem e entram, ressaltando que para fazer atualizações em ontologias, precisamos de um método para extrair uma ontologia a partir de um grande número de interações individuais. O trabalho de Hamasaki *et al.* [11] fornece método para a construção de uma ontologia emergente que leva em conta as relações sociais, considerando os conceitos marcados por cada comunidade como diferentes, mesmo que tenham o mesmo rótulo.

Já o objetivo em Mori *et al.* [12] é extrair as relações subjacentes entre as entidades que são incorporadas em redes sociais. Eles propõem um método que automaticamente extrai rótulos que descrevem as relações entre as entidades da rede. Fundamentalmente, o método faz um agrupamento de pares similares de acordo com seus contextos coletivos em documentos na *Web*. Para extrair os rótulos apropriados de relação, eles extraem rótulos descritivos das relações automaticamente como: filiação, papéis, lugares, relação de todo-parte e relações sociais.

No trabalho de Šimko & Bieliková [9] a criação de cursos *online* adaptativos são a motivação para a extração de semântica, com uma abordagem para o problema de descoberta automática de relacionamento. Eles computam similaridade entre conceitos, que é uma etapa central no processo de criação de um relacionamento. Para isso, propõem um método automático de descoberta de relação entre conceitos para um curso *online* adaptativo. A proposta é fundamentada em duas abordagens baseadas em algoritmos de processamento e análise de grafos subjacentes ao modelo de domínio.

A relação entre redes sociais e a extração de semântica vem sendo discutida na literatura; contudo os trabalhos utilizam abordagens baseadas em "*collaborative tagging*" [8, 11, 24]. Salientamos a importância de abordagens que não sejam exclusivamente baseadas no uso das *tags,* devido a fatores semânticos mais ricos presente em textos que contribuem para uma melhor construção das ontologias; também o uso do conceito de *tags* ainda não é familiar para usuários em fase de letramento digital. Adicionalmente, os trabalhos na literatura têm utilizado extração de dados de redes sociais para fazer análise da rede: *i.e.* para medir força das relações sociais, e descrever relações entre entidades da rede. Contudo, percebemos que estudos que objetivam efetuar extração para a modelagem semântica da linguagem compartilhada na rede ainda são foram realizados. Portanto, objetivamos efetuar a extração

de dados para construir ontologias baseadas na linguagem do usuário, conforme conteúdo introduzido por eles.

3 Ferramentas e Técnicas para a Extração

Nesta seção são apresentadas as ferramentas e técnicas usadas neste trabalho. As ferramentas utilizadas para a extração de termos são o *KEA*[4] [13] e o ExATOlp [14], enquanto que para a descoberta de relações usando técnicas de agrupamento é utilizado o *software CLUTO*[5] [15]. Além disso, é apresentrada uma justificativa da escolha do uso destas e de outras possíveis ferramentas e ambientes, assim como alguns que não são apropriadas aos objetivos deste trabalho.

3.1 Kea

KEA [13] é um *software* para efetuar extração automática de "termos chave" (*keyphrases*). Este recebe "texto bruto" de documentos e extrai um conjunto de possíveis termos relevantes. Utilizando o *KEA* é possível extrair palavras isoladas ou múltiplas palavras que descrevem o assunto de um determinado documento. O algoritmo possui dois estágios: (1) Treinamento: nesta fase é criado um modelo para ajudar na identificação dos termos chave usando um conjunto de documentos de treinamento, onde os termos chave são conhecidos e atribuídos de forma humana. O algoritmo executa o modelo observando valores que podem ser utilizados como exemplos positivos ou negativos na extração; (2) Extração: nesta fase escolhem-se termos a partir de um novo documento, usando o modelo criado na fase de treinamento. A fase de extração tem duas etapas: identificação de candidatos e a seleção dos termos chave. Após a identificação dos candidatos, o algoritmo aplica o modelo para computar a probabilidade de cada candidato ser um termo chave. Então este seleciona os *n* melhores termos conforme parâmetros de entrada. Este *software* é baseado em redes Bayesianas, utilizando algoritmos da ferramenta *WEKA*[6] [16]. Logo, a sua precisão está condicionada principalmente ao conjunto de treinamento efetuado.

3.2 ExATOlp

[4] http://www.nzdl.org/Kea/
[5] http://glaros.dtc.umn.edu/gkhome/views/cluto/
[6] http://www.cs.waikato.ac.nz/ml/weka/

ExATOlp (Extrator Automático de Termos para Ontologias em Língua Portuguesa) [14] é uma ferramenta de *software,* para extrair e selecionar termos significantes e freqüentes de um *corpus* em português anotado linguisticamente. Esta ferramenta é baseada em uma abordagem lingüística e estatística, e ao receber um *corpus* anotado sobre um domínio específico de interesse, extrai automaticamente todos os sintagmas nominais gerando uma lista de termos significativos, e também opcionalmente algumas medidas numéricas. Em outras palavras uma vez que os termos são extraídos, um tratamento estatístico é feito computando freqüências absolutas e relativas de cada um dos termos extraídos.

O processo de extração começa com a anotação lingüística do *corpus*, que é realizada pelo *parser* PALAVRAS [17]. Cada palavra de cada frase é anotada de acordo com sua função sintática, as suas características morfológicas e uma marcação semântica. A ferramenta usa um grupo de heurísticas para refinar o processo de extração, e os sintagmas nominais extraídos são os candidatos a conceito de uma ontologia.

3.3 Cluto

CLUTO (*A Clustering Toolkit*) [15] é um pacote de *software* para agrupamento de conjuntos de dados de alta e baixa dimensão, e para análise das características dos vários agrupamentos. Este pacote oferece três diferentes classes de algoritmos de agrupamento, que operam diretamente tanto no espaço da característica dos objetos ou no espaço de similaridade dos objetos. Estes algoritmos são baseados em paradigmas de partição e aglomeração. Uma característica-chave na maioria dos algoritmos de agrupamento do *CLUTO* é que eles tratam o problema de agrupamento como um processo de otimização, que visa maximizar ou minimizar uma função de critério de agrupamento particular definida globalmente ou localmente durante toda a solução. O *software* oferece um total de sete funções com diferentes critérios e parâmetros que podem ser usados para conduzir os algoritmos de ambos os paradigmas.

3.4 Justificativa das Escolhas

Há diversas propostas na literatura de ferramentas, ambientes e técnicas para a extração de termos e também relações entre estes, que podem auxiliar na construção de ontologias. No entanto, este ainda é um desafio de pesquisa; as soluções já apresentadas na literatura, apesar dos bons resultados, ainda se mostram limitadas e altamente dependentes de contextos bem definidos. Há diversos fatores que guiaram a escolha destas três ferramentas para efetuarmos o experimento neste trabalho: a natureza dos dados, sendo estes um *corpus* de referência; a língua do *corpus*, o português; a fácil disponibilização e utilização das ferramentas, incluindo documentação que possibilite suporte ao experimento, e também seu prestígio e uso na comunidade; além do propósito e abordagem técnica utilizada em cada um.

O KEA é uma ferramenta utilizada na comunidade de mineração de texto. Esta tem a facilidade de poder ser integrada a outras aplicações Java. Adicionalmente, devido a sua

abordagem estritamente estatística, ela extrai rótulos descritivos sem depender de quaisquer características sintáticas do texto, tratando este como uma *"bag of words"*. Em outras palavras não há a necessidade de anotar o texto com informações lingüísticas, podendo assim ser utilizado independente da língua do texto, e sem um difícil e custoso pré-processamento.

A ferramenta ExATOlp é especialmente voltada ao processamento da língua portuguesa, o que parece ser uma boa opção dado o objetivo deste trabalho. Esta utiliza uma abordagem lingüistica voltada a esta língua através da anotação do *corpus*. Uma ferramenta semelhante ao ExATOlp é a OntoLP [18] que implementa entre outras possibilidades de extração, uma abordagem igualmente baseada em informações lingüísticas. A diferença básica dentre estas é que o OntoLP é um *plug-in* para o editor de ontologias Protégé[7], e extrai os termos não somente baseado nos sintagmas nominais mas também possui outros dois métodos como opção: n-gramas e padrões morfossintáticos. Contudo, para obter o padrão de entrada dos dados exigido pela ferramenta OntoLP é necessário um pré-processamento mais árduo do texto original comparado ao ExATOlp, o que dificulta o experimento e está mais propenso a erros. Desta forma, por este experimento não necessitar de outros métodos de extração além do já fornecido pelo ExATOlp, consideramos a utilização desta. Em Lopes & Vieira [19] é descrita uma comparação de métricas do ExATOlp com o OntoLP. Segundo Lopes & Vieira [19:60] os resultados preliminares do uso da ferramenta ExATOlp mostram um desempenho promissor quando comparado a outras ferramentas similares.

Considerando ferramentas que implementam técnicas de agrupamento, além da possibilidade do *CLUTO*, poder-se-ia utilizar o *SenseClusters*[8] ou os algoritmos de agrupamento disponíveis pelo *WEKA*. Como estas ferramentas têm o mesmo propósito, e o *SenseClusters* também utiliza o *CLUTO* como base, optou-se pelo *CLUTO*. Além destes, há diversos ambientes projetados especificamente para suporte à construção semi-automática de ontologias (*e.g. OntoGen*[9]). Uma lista destes ambientes é apresentada e discutida em Baségio [20]. Foi observado que a maioria utiliza abordagem lingüística, que é voltada para outras línguas como a inglesa e francesa, o que os inviabiliza para testes neste trabalho, devido ao contexto em estudo.

[7] http://protege.stanford.edu/
[8] http://senseclusters.sourceforge.net/
[9]http://ontogen.ijs.si/

4 Capturando a Semântica Utilizada na RSI

Os textos utilizados no experimento são ao todo 232 anúncios reais da RSI *VilanaRede* distribuídos entre anúncios de produtos, serviços e idéias. Nestes estão os principais dados que contém o vocabulário compartilhado pelas pessoas participantes desta rede social.

Para a ferramenta KEA foi escolhido um terço de todos estes anúncios de forma aleatória, e os "termos chave" foram indicados de forma manual para o treinamento da ferramenta e a criação do modelo. Foi também utilizada uma lista de *stopwords* em português, que são termos sem sentido que devem ser desconsiderados pela ferramenta durante sua análise. Ao se aplicar o KEA, os resultados de saída são arquivos com os "termos chave" processados para cada arquivo de entrada. Os mesmos arquivos de texto que foram entrada para o KEA, foram anotados utilizando o *parser* PALAVRAS. Os arquivos de saída deste são do formato *TigerXML* [21], que são utilizados como entrada para o ExATOlp. Esta ferramenta retorna três listas como resultado, uma de cada tipo: unigramas, bigramas e trigramas, que estão ordenadas pela relevância do termo. Baseado neste resultado, um procedimento escrito em Java foi desenvolvido, que faz um pós-processamento nestas listas, agrupando os termos pela sua correspondente marcação semântica.

Já a ferramenta *CLUTO* utiliza como entrada uma matriz que armazena os objetos a serem agrupados. Para converter os documentos de texto para esta matriz de formato aceito pelo *CLUTO* utilizamos um *script* em Perl chamado *doc2mat*[10], que também utilizou a mesma lista de *stopwords* usado pelo *KEA*. O *script doc2mat* converte um conjunto de documentos texto em um formato de espaço vetorial utilizado pelo *CLUTO*. Cada linha desta matriz representa um único objeto, neste caso os anúncios com seus comentários, enquanto as suas diferentes colunas correspondem às dimensões dos objetos. Neste experimento utilizamos diversas configurações distintas para analisar e testar os dados. Um dos parâmetros do *CLUTO* é a quantidade de grupos, e o resultado são os grupos formados pelos possíveis termos com medidas estatísticas para cada termo.

Foram analisados os resultados da saída de cada ferramenta isoladamente; após isso, foi contraposto cada resultado de uma ferramenta com a outra utilizando um procedimento em Java desenvolvido neste estudo com o intuito de encontrar e analisar termos que possivelmente se repetiriam nas três. Este procedimento também possibilita contrapor os grupos semânticos processados a partir dos resultados da ExATOlp, com os grupos gerados pelo CLUTO.

[10]http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.html

4.1 Resultados

Kea

Com esta ferramenta foram utilizadas duas abordagens para a extração de termos. Uma a partir de cada documento, extraindo no máximo 7 termos chave de cada anúncio. A segunda abordagem envolveu extrair no máximo 50 termos chave de todos os anúncios juntos em um mesmo documento. Sobre a primeira abordagem, mostramos alguns casos de sucesso da extração, e também alguns casos que se mostraram deficitários. Esta análise é fundamentada no próprio conteúdo dos anúncios utilizados no experimento. A Tabela 1 (lado esquerdo) ilustra casos positivos; cada linha descreve termos chave extraídos de um anúncio específico. Os termos são descritos conforme encontrados, sem correções ortográficas.

Tabela 1: Alguns casos da extração utilizando o *KEA*

Termos chaves extraídos (Positivos)	**Termos chaves extraídos (Negativo)**
Cida, escova, cabelos, cabelo, corte, unissex, química	Ateliê, estilos, cores, ateliê estilos, estilos e cores
Samuca, festa junina, festança, venham, típicas, prestigiar	Rir, saúde, né, levantar, astral, sabe, mundo
Exercícios físicos, hipertensos, praticar, pressão arterial, físico	Rock, morrido, CD, pensei, tinha, sabe
Reeducação alimentar, restritiva, ideal, alimento, gordura quantidade	Ideia, Lina, acho, complementação, valida, comunitarios igrejas
Desfile, ecofashion, feira, solidária, ateliê, estilos economia	
Chinelos, decorados, havaianas, ótimo, preço, peças, confeccionados	

Os casos descritos na Tabela 1 (lado direito) foram considerados de sucesso, uma vez que os termos retornados são expressivos para os anúncios considerados, no sentido de conceitos que retratam o assunto tratado no anúncio. Apenas observando estes termos podemos ter uma idéia do assunto tratado pelo anúncio, e isto pode ser confirmado ao se ler o anúncio.

É importante ressaltar que em grande parte dos casos dos 232 anúncios (90%), a partir dos termos identificados pelo *KEA* conseguimos ter uma idéia do assunto tratado no anúncio. No entanto, em alguns casos, conforme ilustra a Tabela 1 (lado direito) os termos são repetidos ou sem sentido, não sendo boas representações do anúncio, do qual os termos foram extraídos. Observe que há termos como: "pensei", "tinha", "sabe" que não têm o poder de expressar o assunto tratado no anúncio e também não são termos adequados à lista de *stopwords*. Lendo apenas estes termos é difícil obter uma visão do assunto do anúncio. É

também importante salientar que esta ferramenta não retorna apenas substantivos. Verbos também são retornados como termos chave, o que pode possibilitar construir ontologias mais elaboradas, dependendo da metodologia adotada para sua construção. Foi observado que qualitativamente, no geral, os casos negativos não têm o poder de atrapalhar ou influenciar gravemente o resultado da extração.

Considerando a segunda abordagem feita com o *KEA*, podem-se identificar diversas características nos termos extraídos, como termos informais: "oi", "olá", "vc"; e também conceitos importantes utilizados na rede como: "internet", "saúde", "blog", "projeto", "rede", "anúncio" e "Chácara de Orgânicos". Sobre este último termo há diversos anúncios na rede relativos a este assunto, que contém este termo, o que faz todo sentido tal termo ser indicado como relevante.

ExATOlp

Nesta subseção são apresentados os resultados coletados a partir desta ferramenta, e também os resultados do pós-processamento.

Tabela 2: Lista com os 15 termos mais relevantes extraídos pelo ExATOlp

Unigrama			**Bigrama**		
Termo	**FA**	**FR**	**Termo**	**FA**	**FR**
pessoas	78	0.0111620	exercício_físico	19	0.00271895
internet	63	0.0090154	e_mail	18	0.00257584
rede	56	0.0080137	meio_ambiente	14	0.00200343
trabalho	55	0.0078706	Ateliê_Estilos	13	0.00186033
vila	35	0.0050085	reforma_ortografica	13	0.00186033
coisa	31	0.0044361	atividade_física	11	0.00157413
saude	30	0.0042930	Vila_União	10	0.00143102
sites	30	0.0042930	Bom_dia	9	0.00128792
curso	27	0.0038637	mera_título	8	0.00114482
ponto	25	0.0035775	cooperativa_cidarte	7	0.00100172
produtos	25	0.0035775	inclusão_digital	7	0.00100172
vc	24	0.0034344	alimentos_orgânicos	5	0.00071551
alimento	23	0.0032913	Atelie_Tok	5	0.00071551
anúncio	23	0.0032913	Forno_Solar	5	0.00071551
Brasil	23	0.0032913	gordura_corporal	5	0.00071551

Tabela 3: Lista com alguns termos extraídos pelo ExATOlp separados por grupos semânticos

Grupo	Unigrama	Bigrama
Comida	Arroz, alimentícios, açúcar, frutas, pimentão, alimento, óleo, páprica	páprica_doce, alimentos_funcionais, alimento_ideal, arroz_integral, óleo_quente, alimento_estéril, açúcar_café, alimentos_convencionais, alimentos_orgânicos, alimentos_industrializados
Frutas	Uva, bananas, cebolas, abóbora, frutos, macaxeira	ervilhas_lentilhas, uva_integral, cebolas_pequenas caju_nozes, frutos_oleaginosos, amêndoa_gordura
Quantidade	Suficiente, quantidade, sobrecarga, porcentagem, trouxinha, dose	direito_ambiental, sobrecarga_articular, dose_certa
Animal	Cachorro, leque, bichinhos	bichos_peçonhentos, cachorro_quente
Profissão humana	Médico, logistas, governantes, artesãs, educadora, professores, repórter, empregados, padres, professora, agricultor, cabeleireiro, aluna, pesquisadores, enfermeiros, produtores, engenheiro, assistente, nutricionista, pediatra, microempresário, atendentes, educador, agrônomos, farmacêutico	repórter_empresários, boa_professora, educadora_ambiental, costureira_cabeleireira, assistente_administrativo, pesquisador_UNICAMP, agricultor_orgânico, educador_físico, novos_empreendedores
Sem categoria	Novelinho, rasteirinhas, florzinha, nozinhos, minino, lembrancinhas, Olhadinha, noivinha, bracinho, fohlinhas, toquinhas, mural	cooperativa_cidarte, quadradinhos_prontos, novo_quadradinho, excelente_crocheteira, Posto_marajoara

A Tabela 2 mostra uma lista com os 15 termos mais relevantes de unigramas e bigramas extraídos pelo ExATOlp. Esta Tabela apresenta os termos na forma de sintagma no formato original que aparece no *corpus*, assim como sua freqüência absoluta (FA), ou seja, o número de vezes que o sintagma foi corretamente identificado nos textos tratados; e sua freqüência relativa (FR), que é a freqüência absoluta dividida pelo total de termos identificados no *corpus*. Observe que os unigramas com maior freqüência são relativos aos termos: "pessoa", "internet" e "rede". É importante ressaltar que a ferramenta extraiu mais de 1000 termos e há possibilidade de parametrizar esta informação. Outros termos que não constam nesta tabela, mas que vale ressaltar devido a aspectos culturais são: "crochê", "lantejoula" (FA: 7;FR:0.00100172), "quadradinho" (FA:6;FR:0.000858615) e "dízimo" (FA:2;FR:0.000286205). Além dos unigrama e bigramas; os principais trigramas identificados pelo ExATOlp são: "Vila em rede" (FA:23;FR: 0.00329136), "Chácara de Orgânicos" (FA:17;FR: 0.00243274), "aplicação de lantejoulas" e "Santo de pano" (FA:6;FR: 0.000858615).

A Tabela 3 apresenta alguns dos diversos grupos semânticos identificados no pós-processamento a partir dos dados identificados pelo ExATOlp, com unigramas e bigramas. Observam-se alguns resultados como "páprica" e "macaxeira" que são termos culturais de regiões geográficas específicas. Aparecem também diversos nome de profissões encontrados nos anúncios que podem indicar possíveis perfis dos usuários da rede social, assim como o grupo semântico chamado "humano" (não apresentados na Tabela 3) também indica termos neste sentido (sobre característica de perfis). Há alguns resultados para os quais devemos chamar a atenção como: "leque" (Animal), "cachorro quente" (Animal) e "macaxeira" (Fruta). Mesmo o termo "leque" representando um animal, este pode ser utilizado também com outros significados, e é o que ocorre na rede; no anúncio que cita o termo "leque", este é utilizado no sentido de diversas opções, assim como o termo "trouxinha" na rede não tem o sentido de quantidade, e macaxeira não é Fruta. Nem todos os termos extraídos pelo ExATOlp possuem uma marcação semântica. Dentre estes foram observados termos informais utilizados no *VilanaRede*, por exemplo entre os sem categoria na Tabela 3.

Cluto

Em razão da quantidade e diversidade de algoritmos e parâmetros disponibilizados por esta ferramenta, escolhemos 8 configurações distintas entre algoritmos e parâmetros, com 8 grupos de 7 termos para cada configuração, que gerou um total de 64 grupos distintos. A partir destes resultados gerados pelo *CLUTO*, fizemos um filtro manual dos grupos apresentados, identificando e juntando os grupos com termos semelhantes. A seguir filtramos os grupos com os termos que achávamos mais relevantes, baseado no nosso conhecimento do conteúdo dos anúncios. Isto resultou em 24 grupos mais representativos, que foram utilizados durante a interseção de termos entre as ferramentas.

Tabela 4: Alguns casos do agrupamento utilizando o CLUTO

Grupos do *CLUTO* (Positivos)	Grupos do *CLUTO* (Negativo)
Orgânicos alimentos chácara produtos Brazil	Metavendas Cida Brazil metavendasbrazil contato
Natura entrega pronta consultora Avon	Toma freqüência bebida maior currículo alinesantos
Meditação paz cidade prática meditar comprovam	Sociais informa jornal editar rede perfil redes
Curso Unicamp fuxico instituto teste marruda libras	Gostaria Vânia fotos anúncio Neusa casa
Ateliê estilos cores solidária economia seminário femininos feira	
Cristian exercício físico alongamento agita caminhada gordura projeto	

A Tabela 4 (lado esquerdo) apresenta alguns grupos com os termos que foram observados como casos positivos a se considerar. Esta observação foi baseada nos anúncios reais do sistema nos quais estes termos têm certa relação semântica. Diferentemente da primeira abordagem do *KEA*, os termos dos grupos criados pelo *CLUTO* não necessariamente são do mesmo anúncio. É importante ressaltar que em alguns casos, conforme ilustra a Tabela 4 (lado direito), os grupos formados são compostos de termos que muitas vezes não possuem algum tipo de relação como de co-ocorrência. Observando casos da Tabela 4 (lado direito) não encontramos relação entre os termos da mesma linha nos anúncios da rede social. Tivemos dificuldade em identificar e avaliar qualitativamente se os grupos gerados estavam consistentes com a realidade ou não. Dos 24 grupos filtrados, obteve-se aproximadamente 15 grupos de maior qualidade, o que corresponde a aproximadamente 62%.

Um aspecto interessante a se considerar é que o *CLUTO* gerou nome de pessoas em diversos grupos. Fato este que pode indicar que uma pessoa está mais fortemente conectada com um conceito utilizado na rede. Neste experimento também foi feita uma tentativa de agrupamento utilizando como entrada de dados os termos chave produzidos pelo *KEA*; no entanto não obtivemos bons resultados. Acreditamos que isso se deva à pouca quantidade de dados de entrada disponível, que limita a ação dos algoritmos do *CLUTO*.

Conforme a Tabela 4 (lado direito), estes termos foram considerados negativos do agrupamento utilizando o *CLUTO* uma vez que apresentam palavras que, tendo conhecimento dos anúncios na rede social, pode-se observar que não faz sentido uma palavra com a outra (*i.e.* não traz riqueza na relação entre os termos).

Interseção entre Termos e Grupos

Para verificar quais termos se repetem no *KEA*, ExATOlp e *CLUTO*, assim como verificar se os agrupamentos do *CLUTO* têm alguma interseção com os agrupamentos semânticos processados a partir do ExATOlp, foi feito um processamento utilizando o resultado destas três ferramentas. A Tabela 5 mostra os termos identificados igualmente pelas três ferramentas. Note que alguns termos como: "vila" e "produtos" são termos indicados com alto FR pelo ExATOlp, mas por exemplo o termo "pessoa" que possui a maior FR pelo ExATOlp não se encontra na lista.

Tabela 5: Termos igualmente extraídos pelas 3 ferramentas e alguns termos que estão na interseção de resultados de duas ferramentas

Interseção	Termos	Interseção	Termos
Interseção entre as 3 ferramentas	Vânia, casa, cidade, vila, anúncio, curso, orçamento, contato, maior, freqüência, produtos, ponto, oficina, linha crochê, artesanato, fuxico, tecido, rede, boneca, fio, rock, projeto, alongamento, bicicleta, gordura, caminhada, fotos, rosto	***KEA-CLUTO***	Femininos, pano, alimentos, artesanal, desfiar, almofada, preço, orgânicos, lindo, divulgar, bebida, multimídia, comunitário
KEA-ExATOlp	Ovo, queijo, futebol, xadrez, mundo, rasteirinhas, minhocário, comunicação, chocolate, bordados, educação, toquinhas, decoupage, moradores, inclusão	***ExATOlp-CLUTO***	Quartas, jornal, seminário, blog, exercício, hora, feira

A Tabela 5 também apresenta alguns termos encontrados que estão na interseção de apenas 2 ferramentas. Importante ressaltar que a grande maioria das interseções aconteceram entre termos do *KEA* e do ExATOlp. Com relação à interseção entre grupos do *CLUTO* e do ExATOlp, referente ao agrupamento utilizando as marcações semânticas, constatamos que apenas os termos "linha" e "fio" tiveram interseção entre um grupo do ExATOlp, chamado de roupa, com um grupo do *CLUTO*.

4.2 Discussão

Devido à complexidade e diversidade da linguagem humana, assim como das possibilidades de representação mental do conhecimento desenvolvidas pelo ser humano, apenas capturar termos e possíveis relações não parece suficiente para expressar fielmente a semântica

compartilhada na rede social. Uma ontologia está em um nível mais alto de abstração, e certamente não representará completamente a linguagem natural. Entretanto, o esforço para melhor capturar e representar os significados compartilhados pelas pessoas organizadas em comunidades *online* pode influenciar não apenas informando buscas mais adequadas ao usuário, mas também diversos outros elementos que levem em consideração os significados utilizados pelas pessoas. Identificar os possíveis conceitos e suas relações representa um primeiro passo neste objetivo maior.

O cenário de RSI é um contexto ainda não explorado que necessita investigação aplicando técnicas e métodos presentes na literatura para observar e verificar o quanto são apropriadas. No contexto do experimento desenvolvido neste trabalho, foi observado que a abordagem mais apropriada para melhor analisar os dados é verificar tanto os dados capturados por anúncio individualmente, quanto todos os dados juntos desconsiderando a individualidade do conteúdo. Isto permite verificar os dados de duas perspectivas diferentes.

Com os termos extraídos do *KEA* sobre cada anúncio pode-se ter uma idéia dos assuntos tratados na rede. Já a abordagem utilizada pelo ExATOlp fornece uma visão geral de todos os dados da rede. Outro fator a considerar é que o ExATOlp extrai apenas substantivos, enquanto o *KEA* pode considerar também verbos, que em algumas situações, na nossa percepção, podem ajudar na construção de ontologias melhor elaboradas. O *KEA* utiliza uma abordagem estatística, que tem a vantagem de se adaptar mais facilmente independente do domínio e língua do *corpus*. O ExATOlp utiliza uma abordagem lingüística, na qual é necessário anotar o *corpus*. Isto gera dificuldades e pode ser um entrave crítico ao se utilizar esta ferramenta para muitos conteúdos.

Em contrapartida, o ExATOlp oferece melhores resultados para observar os dados, com medidas estatísticas: freqüência absoluta e relativa, é mais organizado e de melhor entendimento. Com isso é possível observar melhor os termos mais utilizados na rede. A partir do pós-processamento efetuado, os termos organizados pela categoria semântica mostram uma perspectiva interessante dos dados, uma vez que permite ao engenheiro de ontologia conhecer os termos semanticamente relacionados. Dentre as vantagens do ExATOlp tem-se: possuir as medidas dos termos, o que não é gerado pelo *KEA,* incluir a marcação semântica para cada termo, não necessitar de uma lista de *stopwords* e nem de treinamento *apriori*, diferentemente do *KEA*. Analisando os resultados do *CLUTO*, pode-se observar uma qualidade mais baixa dos resultados extraídos comparado ao KEA e ExATOlp, e também baseado no conhecimento do conteúdo dos anúncios. Isto provalmente se deva à natureza dos dados. Consideramos que melhores resultados com o *CLUTO* poderiam ser obtidos caso uma massa maior de dados fosse considerada.

Os termos extraídos através dos dados reais da rede social com as ferramentas são insumos importantes na criação de ontologias. Observar e analisar os termos que se repetem nas três ferramentas com mais atenção, conforme ilustra a Tabela 5 é importante, pois estes podem ser uma boa indicação de possíveis conceitos que não devem ser desconsiderados na modelagem da ontologia. Além destes, é possível considerar diversos outros elementos que podem gerar pequenas partes de ontologias. Analisar os grupos identificados pelo *CLUTO*, assim como os termos extraídos de cada anúncio pelo *KEA* pode contribuir para modelar partes de ontologias, uma vez que estes indicam termos que possivelmente têm alguma

relação semântica. Como exemplo observe-se na Tabela 4 (lado esquerdo); os termos "exercício físico", "alongamento", "caminhada" e "gordura", que têm uma relação semântica no mundo real. Assim como os termos bigramas extraídos pelo ExATOlp agrupados por *tags* podem indicar relações genérico-específica; *e.g.* considere os termos retirados da Tabela 3: "alimentos funcionais", "alimento ideal", "alimento estéril", "alimentos convencionais", "alimentos orgânicos". O termo alimento neste exemplo pode ser uma classe mais genérica, enquanto estes tipos de alimentos são conceitos mais específicos relacionados a alimento. Os termos informais presentes na classe sem categoria do ExATOlp são importantes de serem observados, pois expressam palavras utilizadas no contexto daquela rede, que não constam em possíveis dicionários formais. Observe-se na Tabela 3 que a maioria destes termos está no diminutivo, o que representa a maneira como as pessoas envolvidas se expressam no dia a dia. O diminutivo na maioria dos casos não esta associado a medida de tamanho, mas sim a linguagem coloquial para expressar algo carinhosamente ou o corriqueiro.

Não apenas as ferramentas e técnicas isoladamente são importantes, mas também como estas são utilizadas e organizadas para o engenheiro de ontologias tomar decisões referentes à modelagem das ontologias, com base em diversos resultados e perspectivas. Consideramos bons os resultados tanto do *KEA* quanto do ExATOlp dado o contexto em estudo. As estratégias desenvolvidas para intercalar um resultado com outro também devem ser consideradas. Considera-se uma boa estratégia utilizar o *KEA* junto com o ExATOlp e os devidos pós-processamento descritos, que pode levar a resultados satisfatórios, sob o ponto de vista do contexto de RSI. Utilizando desta, o engenheiro terá a oportunidade de observar os termos identificados de diversas perspectivas.

As ferramentas mostraram-se importantes para o processo de identificação de semântica, e podem ajudar na criação de ontologias que representem os significados compartilhados na rede social. Todavia é importante ressaltar que é necessário efetuar uma filtragem com intervenção humana, de forma a eliminar problemas e falsos conceitos. Outras maneiras de observar os dados poderiam ser obtidas através de outros algoritmos. Dentre as outras abordagens para analisar os dados, uma seria tentar descobrir possíveis regras de associação a partir dos dados, através do algoritmo Apriori do *WEKA*. Tais regras poderiam identificar que possivelmente um determinado termo só ocorre quando outro termo ocorre, ou só ocorre com determinado usuário. Isto poderia contribuir para informações adicionais que podem guiar o engenheiro na modelagem.

Outra possibilidade para agrupamento em detrimento do *CLUTO* seria utilizar outros algoritmos, *e.g. SenseClusters* em sua implementação de *Latent Semantic Indexing,* assim como o *WEKA* com o algoritmo de k-médias para observar a diferença de agrupamento com o *CLUTO*. Utilizar técnicas conforme descrito por Deepak [22] também pode ser relevante, assim como explorar melhor as questões pertinentes as relações sociais na rede social conforme observado por Mika [8]. Propostas que visam suportar o processo de construção de ontologias visualmente assim como descrito em [23] são não menos importante para melhorar e facilitar a identificação e relação entre conceitos.

Este trabalho não visou mostrar exatamente todos os passos envolvidos na construção de ontologias automaticamente, dada a complexidade do processo. Uma fase importante deste processo é a captura dos conceitos, que foi foco desta investigação para o contexto considerado. Objetivou-se alcançar uma seleção de termos significativos, com o intuito de verificar a viabilidade da abordagem em RSI. Embora não seja menos relevante apontar como se pretende alcançar as etapas subseqüentes para se obter ontologias completas, este estudo está fora do escopo deste trabalho, mas são caminhos de investigação interessantes a serem perseguidos. No futuro é importante efetuar experimentos de como tais conceitos e relações simples, conforme apontado neste trabalho podem ser transformados em ontologias mais bem elaboradas e complexas. Além disso, é também importante observar todo processo incluindo o uso das ontologias em aplicações específicas (*e.g.* busca baseada em ontologia) na RSI.

Sobre este tópico, estudos com usuários reais e conclusões descritas em [2] mostram que busca semântica pode auxiliar usuários em fase de letramento digital a acessarem informação em RSI. A construção de ontologias é passo fundamental para alcançar o objetivo de um mecanismo de busca que possibilite resultados que façam sentido às pessoas, e ajude no processo de acesso ao conhecimento. Defendemos que o resultado de busca em RSI baseado em ontologia pode ser melhorado quando o processo de construção é baseado nos dados da rede, uma vez que esta é uma maneira de representar os significados compartilhados pelas pessoas na RSI. Para tal objetivo, o processo de construção das ontologias poderia ser facilitado e melhorado.

O mesmo estudo descrito em [2] mostra limitações de mecanismos de busca tradicionais baseados apenas em comparações sintáticas em auxiliar usuários não letrados digitalmente a acessarem os conteúdos disponíveis. Isso reforça e justifica a importância do processo investigado neste trabalho, que parte de dados da RSI para ontologias, para o desenvolvimento de mecanismos de busca semânticos que considerem o linguajar das pessoas da rede. Ou seja, existem experimentos na literatura que comprovam com dados empíricos os problemas de mecanismos de busca sintáticos que motivam a necessidade de uma busca melhor e mais adequada ao público alvo. O desenvolvimento de ontologias para este fim é fator fundamental.

Por fim, mesmo em um cenário não limitado a um contexto ou domínio específico do conhecimento (*i.e.* corpus de referência), o que pode dificultar a ação das ferramentas, e considerando as limitações destas, ressaltamos que a utilização das ferramentas pode contribuir para o objetivo deste trabalho, sendo um primeiro passo para uma representação mais adequada da semântica em RSI. Conseqüentemente este estudo pode contribuir para o desenvolvimento de mecanismos de busca mais apropriados a RSI, onde se julga importante respeitar as limitações e diversidade das pessoas enquanto construtoras de significados.

5 Conclusão

Mais importante do que a quantidade de informação é a sua relevância para o usuário. Mecanismos de busca em RSO poderiam ser mais adequados à linguagem cotidiana das

pessoas, e aos significados que elas compartilham em comunidade. Em RSI isto pode ser fator determinante para o usuário obter êxito ou não ao tentar acessar informação. Representar a semântica compartilhada na rede passa obrigatoriamente por descobrir conceitos utilizados e suas relações semânticas. Este trabalho apresentou possíveis ferramentas e técnicas que podem auxiliar nesta tarefa em um contexto de RSI. Os resultados aplicando as ferramentas e técnicas em dados reais se mostraram promissores no sentido de auxiliar na construção de ontologias que representem os significados utilizados na rede social, possibilitando, portanto, a elaboração de mecanimos de busca mais adequados, principalmente em comparação com mecanismos tradicionais baseados em comparação léxico-sintáticas.

Como trabalhos futuros, prevemos novos testes com outros *softwares* apontados na discussão, assim como pesquisa por métodos que visam avaliar com usuários da rede social e engenheiros de ontologias quais dos termos extraídos são mais apropriados e como podem auxiliar na construção de ontologias mais adequadas em contextos de diversidade. Investigações que permitam estratégias para melhor visualizar e integrar os dados em um processo definido, conforme descrito em [23] podem gerar maneiras mais oportunas e práticas para filtrar a informação. Uma possível integração destas estratégias com a abordagem apontada neste trabalho em um ambiente de *software* também pode ser considerada em trabalhos futuros.

Agradecimentos

Este trabalho é financiado pela Microsoft *Research* – FAPESP Instituto para Pesquisas em Tecnologia da Informação (processo nro. 2007/54564-1). Os autores também agradecem a Lucelene Lopes, Mírian Bruckschen e Renata Vieira da Pontifícia Universidade do Rio Grande do Sul (PUCRS), pela atenção e ajuda para anotar o corpus e com a ferramenta ExATOlp.

Referências Bibliográficas

1. NERIS, V. P. A., ALMEIDA, L. D., MIRANDA, L. C., HAYASHI, E., BARANAUSKAS, M. C. C., 2009. **Towards a Socially-constructed Meaning for Inclusive Social Network Systems.** In Proc. of Int. Conf on Informatics and Semiotics in Organisations. pp. 247-254.
2. REIS, J.C., BONACIN, R., BARANAUSKAS, M.C.C. 2010. **New Perspectives for Search in Social Networks: A Challenge for Inclusion.** In 12th International Conference on Enterprise Information Systems (ICEIS), Funchal, pp. 53-62.
3. BERNERS-LEE, T., HENDLER, J., LASSILA, O., 2001. **The Semantic Web**, Scientific American.

4. USCHOLD, M., 2003. **Where are the semantics in the semantic web?** American Association for Artificial Intelligence. AI Magazine. Volume 24- Issue 3. pp. 25-36. Menlo Park, CA, USA.
5. SILVA S.R.P. AND R. PEREIRA. 2008. **Aspectos da Interação Humano-Computador na Web Social,** in VIII Simpósio Brasileiro de Fatores Humanos Sistemas Computacionais. ACM Vol. 378. pp. 350-351.
6. GRUBER, T. R., 1993. **A translation approach to portable ontologies**. Knowledge Acquisition, v.5, n.2.
7. BUITELAAR, P., CIMIANO, P., MAGNINI, B. 2005. **Ontology learning from text: An overview.** In Ontology Learning from Text: Methods, Evaluation and Applications, v. 123 of Frontiers in Artificial Intelligence and Apllications. IOS Press.
8. MIKA, P. 2005. **Ontologies are us: A unified model of social networks and semantics**. In Proc. of the 4th Inter. Semantic Web Conf. LNCS 3729, Springer-Verlag.
9. ŠIMKO, M AND BIELIKOVÁ, M. 2009. **Automatic Concept Relationships Discovery for an Adaptive E-course.** In Proc of the 2nd International Conference on Educational Data Mining. Cordoba, Spain. pp. 171-179.
10. ALEMA-MEZA, B. *et al.* 2006. **Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection**. In Proc. of the 15th International World Wide Web Conference.
11. HAMASAKI, M., MATSUO, Y., NISHIMURA, T., TAKEA, H. 2008. **Ontology Extraction by Collaborative Tagging with Social Networking**. 17th Inter. WWW Conference.
12. MORI, J., ISHIZUKA, M., MATSUO, Y. 2007. **Extracting Keyphrases to Represent Relations in Social Networks from Web**. In Proc. of the 20th Inter. joint Conf. on Artificial Intelligence table of contents. Hyderabad, India. pp. 2820-2825.
13. MEDELYAN, O. AND I. WITTEN, H. 2008. **Domain-independent automatic keyphrase indexing with small training sets**. Journal of the American Society for Information Science and Technology. V. 59, I. 7, pp. 1026- 1040.
14. LOPES, L., FERNANDES, P., VIEIRA, R., FEDRIZZI, G. 2009. **ExATOlp: An Automatic Tool for Term Extraction from Portuguese Language Corpora**. In Proc. of the 4th Language and Technology Conference. pp. 427-431.
15. KARYPIS, G. 2002**. CLUTO: a clustering toolkit.** Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. Available at http://glaros.dtc.umn.edu/gkhome/views/cluto/
16. WITTEN, I. H. AND FRANK, E. 2005. **Data Mining: Practical machine learning tools and techniques**. 2nd Edition, Morgan Kaufmann, San Francisco.
17. BICK, E. 2000. **The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.** PhD thesis, Arhus University.
18. RIBEIRO, L.C. 2008. **OntoLP: Construçãoo semi-automática de ontologias a partir de textos da língua portuguesa.** Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Rio dos Sinos.
19. LOPES, L.; VIEIRA, R. 2009. **ExATOlp: Extrator Automático de Termos para Ontologias em Língua Portuguesa.** Relatório Técnico - PUCRS.

20. BASÉGIO, T.L. 2006**. Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil.** Dissertação. Faculdade de Informática, PUCRS.
21. KONIG, E.; LEZIUS, W. 2003. **The TIGER language - A Description Language for Syntax Graphs.** Formal Definition. Technical report, University of Stuttgar
22. DEEPAK, P., RAO, D., Khemani, D. 2006. **Building Clusters of Related Words: An Unsupervised Approach**. PRICAI 2006: Trends in Artificial Intelligence. Lecture Notes in Computer Science – Springer Berlin. Vol. 4099/2006, pp. 474-483.
23. HERNANDES, E. M., SANDE, D., Fabbri, S. 2010. **ONTOP: A process to support Ontology Conceptualization.** In 12th International Conference on Enterprise Information Systems (ICEIS), pp. 58-65, Funchal, June
24. WU, X., ZHANG, L., YU, Y. 2006. **Exploring social annotations for the Semantic Web.** In Proc of the 15th International Conference on the World Wide Web.
25. EGGER, M., FISCHBACH, K., GLOOR, P., LANG, A, SPRENGER, M. 2009. **Deriving taxonomies from Automatic Analysis of Group Membership Structure in Large Social Networks.** In: Lecture Notes in Informatics, vol 154, In Proc of Informatik, Lübeck.