Extração de relações semânticas via análise de correlação de termos em documentos

Sergio William Botero[1], **Ivan L. M. Ricarte**[1]

[1] Departamento de Engenharia de Computação e Automação Industrial
Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 6101 – 13083-970 – Campinas, SP, – Brazil

sergio.botero@gmail.com, ricarte@fee.unicamp.br

Abstract. *Ontologies are important to organize and describe information, but are hard to create and maintain, which motivates the development of tools to help in this task. This article presents a strategy to extract, from a corpora of documents in a given domain, semantic elements expressing proximity relations between terms and concepts to help the construction of domain ontologies. The technique presented here, ACT, is based on linguistic processing, machine learning, and biclustering. Results show that concepts obtained by ACT are at least as good as those from similar techniques, such as LSI and NMF. In relation to those techniques, it additionally has the advantage of allowing the supervision by a domain expert.*

Resumo. *Ontologias são importantes para organizar e descrever informação, mas são difíceis de criar e manter, o que motiva o desenvolvimento de ferramentas de auxílio nessa tarefa. Este artigo apresenta uma estratégia para extrair, de documentos de uma área de conhecimento, elementos semânticos que expressam relações de proximidade de termos a conceitos para auxiliar a construção de ontologias de domínio. A técnica proposta, ACT, é baseada em processamento linguístico, aprendizado de máquina e biagrupamento. Resultados mostram que os conceitos obtidos por ACT são pelo menos tão bons como aqueles obtidos por técnicas similares, como LSI e NMF. Em relação a essas, ela apresenta a vantagem de permitir a supervisão por um especialista de domínio.*

1. Introdução

Ontologias de diversas áreas do conhecimento têm sido utilizadas para apoiar aplicações como busca, processamento de linguagem natural e desenvolvimento de software. Na área de recuperação da informação, elas agregam maior valor semântico a documentos e contribuem para a obtenção de resultados mais efetivos.

A construção de uma ontologia envolve a identificação de diversas entidades ontológicas e seus relacionamentos. Tal tarefa é complexa, custosa e sujeita a erros humanos, o que motiva a utilização de técnicas automáticas ou semi-automáticas. Essas técnicas podem ser baseadas em métodos linguísticos [Hearst, 1998], em métodos estatísticos [Dumais et al., 1995], ou em métodos híbridos, a abordagem mais indicada segundo Gonzalez and Lima [2003].

Métodos estatísticos empregam técnicas de análise da distribuição de termos em documentos para a composição de uma matriz D que relaciona termos a documentos

e a subsequente decomposição matricial para descoberta de estruturas latentes nessas relações. Latent Semantic Indexing (LSI) e Non-Negative Matrix Factorization (NMF) são exemplos dessa abordagem.

Técnicas baseadas em LSI [Dumais et al., 1995] reduzem a dimensão dessa matriz D por meio de uma aproximação baseada na redução no seu posto, obtida pela decomposição de seus valores singulares em três matrizes. Para uma matriz D relacionando N documentos a L termos, $D \approx USV^T$. As matrizes $U_{N\times R}$ e $V_{R\times L}$ constituem bases vetoriais ortonormais que relacionam termos a conceitos e conceitos a documentos, respectivamente. A matriz $S_{R\times R}$ representa a importância de cada conceito para a composição da matriz D, sendo R o posto utilizado na aproximação.

Técnicas que utilizam LSI apresentam dois problemas. Há uma dificuldade para interpretar os valores presentes nos vetores de U e V, devido à mistura de valores positivos e negativos. Outra limitação é imposta pela restrição de ortogonalidade entre os vetores, que resulta em conceitos que não representam corretamente aqueles abordados pelos documentos.

Shahnaz et al. [2006] e Lee and Seung [1999] também propõem a decomposição da matriz D em componentes principais, mas utilizam a fatoração de matriz não-negativa (NMF). Esta técnica resolve os problemas mencionados, pois as matrizes resultantes da fatoração sempre apresentam valores não negativos e não são necessariamente ortogonais, com $D \approx WH^T$. A matriz $W_{N\times R}$ representa a relação entre termos e conceitos; $H_{R\times L}$, a relação entre conceitos e documentos. Entretanto, NMF não permite que o processo de extração de conceitos seja supervisionado e faz-se necessário especificar antecipadamente o número R de conceitos a serem extraídos.

Existem, ainda, os métodos estatísticos baseados em técnicas de agrupamento. Horng et al. [2005] propõem um método de agrupamento fuzzy hierárquico para a determinação de relações semânticas e Fortuna et al. [2006] utilizam LSI e técnica de agrupamento k-means para a construção semi-automática de ontologias.

O presente artigo apresenta uma técnica híbrida para a extração semi-automática de relações semânticas de proximidade entre termos e conceitos. Os métodos lingüísticos realizam a identificação dos termos a partir de um conjunto de documentos e os métodos estatísticos realizam a análise de correlação de termos (ACT) para a identificação de conceitos e suas relações. A nova técnica mostra-se promissora resolvendo os problemas de interpretabilidade de LSI e a falta de supervisão de NMF.

2. Extração de relações semânticas

O modelo ontológico adotado neste trabalho segue o modelo relacional fuzzy [Pereira et al., 2009], que adota dois níveis de abstração para descrever o vocabulário e a organização de um universo do discurso. Um nível representa as entidades mais concretas, denominados de termos; o outro, representa entidades mais abstratas, denominados conceitos. Associações fuzzy descrevem o grau de proximidade entre termos e conceitos. A construção automática de relações semânticas nesse modelo envolve a determinação dos termos, dos conceitos e suas associações.

A Figura 1 apresenta os procedimentos envolvidos nessa atividade. O primeiro estágio, *Pré-Processamento*, utiliza métodos lingüísticos para identificar a lista de termos.

O segundo e terceiro estágios utilizam métodos estatísticos para determinar as matrizes do modelo, os conceitos e as relações. O resultado final é um conjunto de elementos semânticos que serve de base para a construção de ontologias de domínio.

Métodos lingüísticos são utilizados para identificar termos que compõem o vocabulário da ontologia. As palavras contidas nos documentos são extraídas por um analisador léxico e categorizadas de acordo com a sua função morfossintática. Essa categorização é realizada por meio de um processo de rotulagem POS (*part-of-speech*). Neste trabalho, a rotulagem é realizada pela base lexical WordNet[1]. As palavras obtidas são filtradas para obter uma lista de termos da ontologia.

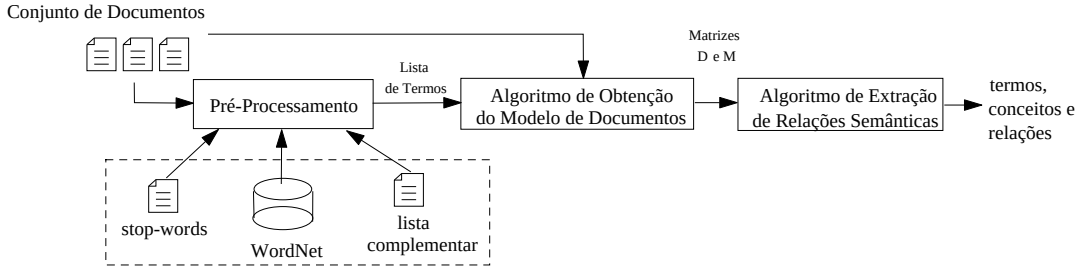

Figura 1. Fluxo de processamento dos dados

A filtragem remove *stop-words* e palavras que não sejam substantivos ou adjetivos. Adicionalmente, a filtragem remove palavras não relevantes para a descrição dos documentos. A seleção dos termos relevantes é baseada no valor de entropia proposto por Velardi et al. [2001]. Uma vez computada a entropia e_t de cada termo, serão selecionados aqueles com $\alpha \leq e_t \leq \beta$, sendo $0 \leq \alpha < \beta \leq 1$. O objetivo desse filtro é remover termos muito comuns ($e_t > \beta$) e termos que ocorrem em poucos documentos ($e_t < \alpha$).

A utilização do WordNet para a identificação de termos permite a classificação automática das palavras, mas pode ignorar palavras específicas de um domínio que sejam importantes para a descrição dos documentos. Dessa forma, o filtro é modificado para permitir a inclusão dessas palavras por meio de uma lista complementar.

3. Análise de correlação de termos (ACT)

A matriz D relaciona termos a documentos, sendo que cada elemento $d_{i,j}$ é calculado pela medida normalizada da frequência do termo pela frequência inversa de documentos [Salton and Buckley, 1988]. No entanto, para estabelecer a organização de termos em um conceito, é preciso computar a correlação entre termos.

Na linguagem escrita, termos relativos a um determinado conceito costumam aparecer de forma correlata nos documentos [Peat and Willett, 1991]. Assim, a correlação entre termos obtidos de um conjunto de documentos de um mesmo domínio, expressa por $M = D^T D$, define uma métrica de similaridade que pode ser utilizada por algoritmos de aprendizado não-supervisionado para a identificação de conceitos. Como D é normalizada, a similaridade é um valor entre 0 (nenhuma correlação) e 1 (máxima correlação).

A matriz M apresentada na Figura 2 ilustra a correlação entre seis termos extraídos de um conjunto de documentos da área de inteligência computacional. Termos

[1] `wordnet.princeton.edu`

semelhantes são agrupados e formam um protótipo de conceito. Protótipos são representados por sub-matrizes de M, não necessariamente contíguas, que reúnem termos com alto índice de correlação. No exemplo, o algoritmo identifica dois grupos, um formado pelos termos *genetic*, *algorithm* e *operator*, e outro, pelos termos *machine*, *learning* e *application*. A determinação dos protótipos de conceito serve de base para o cálculo da associação fuzzy entre termos e conceitos.

$M =$

	genetic	***algorithm***	***operator***	***machine***	***learning***	***application***
genetic	1.000	0.732	0.537	0.003	0.128	0
algorithm	0.732	1.000	0.496	0.010	0.156	0.046
operator	0.537	0.496	1.000	0.183	0.229	0.159
machine	0.003	0.010	0.183	1.000	0.902	0.296
learning	0.128	0.156	0.229	0.902	1.000	0.307
application	0	0.046	0.159	0.296	0.307	1.000

Figura 2. Matriz de correlação de termos

O algoritmo de identificação de protótipo para os conceitos deve agrupar os termos com maior correlação e fornecer métricas para determinar o número adequado de sub-matrizes. Para a identificação das sub-matrizes é definida uma função objetivo, apresentada na seção 3.1, que avalia a correlação entre os termos, e um algoritmo que minimiza essa função. O algoritmo utiliza o método de biagrupamento, que demonstrou ser eficaz na identificação de sub-matrizes em análises de expressões gênicas [Cheng and Church, 2000]. Uma estratégia iterativa e interativa, descrita na seção 3.2, é adotada para a identificação do número de sub-matrizes.

3.1. Função objetivo e algoritmo para agrupamento de termos correlatos

O método de biagrupamento requer a especificação de uma função objetivo F que avalia a qualidade do agrupamento. Seja T o conjunto de todos os termos e $T_c \subseteq T$ o conjunto de termos que definem um conceito c. A função objetivo deve ser tal que $F(T_c) = \Phi(T_c) + \eta\Theta(T_c)$, onde Φ é uma função que avalia a coesão entre os termos e Θ é uma função que avalia a cobertura do conceito. O parâmetro η é um fator de ponderação entre Φ e Θ.

A função de coesão, definida como $\Phi(T_c) = \sum_{i \in T_c} \sum_{j \in T_c} (1 - m_{ij})^2$, avalia a distância entre os termos do agrupamento, sendo $m_{i,j}$ a correlação entre os termos i e j e $(1 - m_{i,j})$ uma medida da distância entre eles. A função de cobertura, definida como $\Theta(T_c) = (L - |T_c|)$, avalia o tamanho do agrupamento. As duas funções devem ser minimizadas para obter conceitos com grande cobertura e coesão entre seus termos.

O algoritmo de otimização da função objetivo é baseado na estratégia de biagrupamento de Cheng and Church [2000]. Ele é iniciado com uma proposta de agrupamento e realiza melhoras por meio de três operações: a inclusão de termos não presentes no agrupamento, a remoção de termos que pertencem ao agrupamento e, se alguma das operações anteriores não for aplicável, pela troca de um termo do agrupamentos por outro qualquer. Uma operação melhora o agrupamento se sua aplicação reduz o valor da função objetivo. O algoritmo termina quando a aplicação das operações não melhora o agrupamento.

3.2. Determinação iterativa de conceitos

A definição da quantidade de conceitos é guiada pela função objetivo. O algoritmo identifica, em ordem decrescente, os mínimos mais pronunciados dessa função. Para isso, o algoritmo deve ser capaz de identificar o mínimo global e ao mesmo tempo evitar os mínimos já descobertos.

Para evitar a ocorrência de conceitos já determinados, uma componente de contexto Γ foi adicionada à função objetivo, $F_{\text{ctx}} = \Phi(T_c) + \eta\Theta(T_c) + \tau\Gamma(T_c)$. Essa componente, $\Gamma(T_c, \bigcup_{i=1}^{N_C} C_i) = \sum_{i=1}^{N_C} \text{sim}(T_c, C_i)$, elimina os pontos de mínimo já descobertos por meio de uma função cuja característica é apresentar pontos de máximo nos mesmos locais dos de mínimo da função F. O parâmetro τ controla a influência da função contexto e a similaridade entre conceitos é expressa por $\text{sim}(T_c, C_i) = (1/|T_c|) \sum_{j \in T_c} \max_l(\bigcup_{l \in C_i} m_{jl})$ que retorna um valor entre 0, não sobrepostos, e 1, quando os conceitos são iguais.

O último passo é o refinamento do conceito, necessário para evitar eventuais distorções causadas pela introdução da função Γ. Para isso utiliza-se novamente a função F para uma busca local, na qual o ponto de partida é o conceito descoberto anteriormente.

A métrica utilizada para determinar o número adequado de conceitos é baseada na taxa de sobreposição entre o conceito extraído e aqueles presentes no contexto. A Figura 3 ilustra uma situação na qual existem N_c conceitos no contexto. A identificação do próximo conceito, T_c, é sucedida pela avaliação de todas as similaridades entre este e os demais conceitos no contexto. A taxa com o maior índice de sobreposição é escolhida; no exemplo, a maior taxa foi obtida com o conceito C_i. Se essa taxa for menor que um limiar γ, T_c é adicionado ao contexto e o usuário continua com o processo de extração. Caso contrário, o usuário cessa o processo de extração, pois não há outros conceitos presentes.

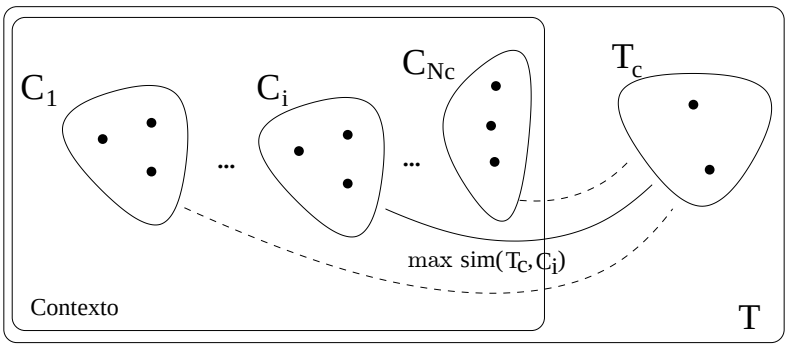

Figura 3. Sobreposição entre conceitos

A iteratividade para a determinação dos conceitos permite a supervisão do processo, introduz flexibilidade para a definição de métricas para determinar o número adequado de conceitos e reduz a complexidade computacional se comparada a técnicas por batelada. Adicionalmente, possibilita a inclusão de novos conceitos pelo especialista que interage com o sistema. Nesse caso, o usuário especifica o conceito por meio da seleção dos termos e pode adicioná-los ao contexto.

Uma visão global dessa estratégia iterativa é ilustrada na Figura 4. Inicialmente, o usuário (um especialista do domínio) escolhe os parâmetros do algoritmo e solicita a

extração de um protótipo de conceito. O sistema identifica, refina e apresenta o protótipo ao usuário, que pode aprová-lo ou não. Se o conceito for aprovado, o usuário atribui-lhe um nome. Por fim, o sistema calcula as associações fuzzy entre os termos e o protótipo de conceito, como descrito a seguir, e armazena o conceito na base de dados.

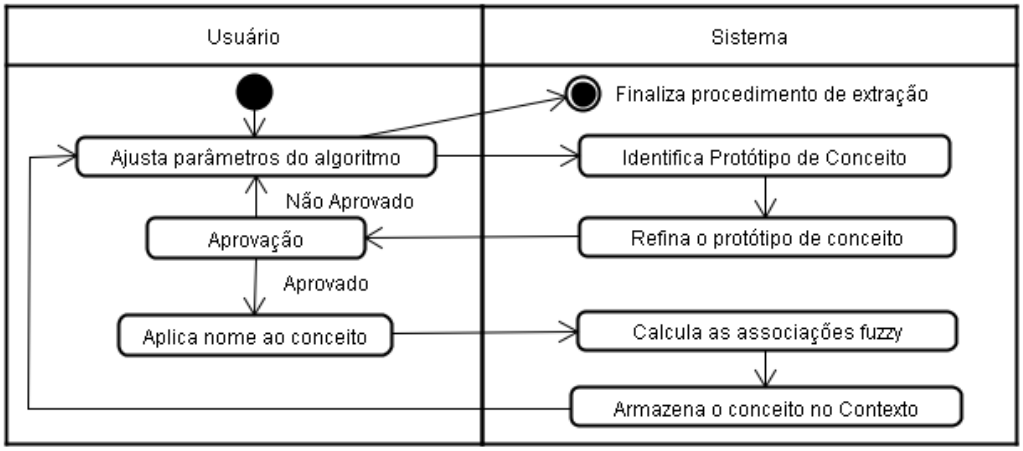

Figura 4. Fluxo de Interação Usuário-Sistema

O grau de associação fuzzy entre termos e conceitos é calculado com base nos protótipos de conceitos determinados pelo algoritmo de agrupamento. Os termos formadores de um protótipo são utilizados para compor uma base vetorial, expressa na forma de uma matriz C na qual as colunas representam os termos que formam o protótipo e as linhas representam o conjunto de documentos. Essa matriz é decomposta utilizando-se a decomposição por valores singulares, $C = USV^T$. A matriz U será a nova base vetorial que define o conceito. Nessa base vetorial os vetores são normais, ortogonais e ordenados pelos valores singulares da matriz S. Seja um termo genérico definido pelo vetor $\overrightarrow{t_i}$; define-se a projeção de $\overrightarrow{t_i}$ em $\overrightarrow{u_1}$, o vetor com o maior valor singular associado, como sendo $\widehat{t_{i,u}} = \text{proj}_{u_1}(\overrightarrow{t_i})$. O valor da associação fuzzy entre o termo $\overrightarrow{t_i}$ e o conceito é dado pelo cosseno do ângulo formado pelos vetores $\overrightarrow{t_i}$ e $\widehat{t_{i,u}}$.

4. Resultados

A técnica ACT foi avaliada segundo dois critérios. O primeiro simplesmente avaliou a habilidade de encontrar um número adequado de conceitos e de agrupar termos de forma coerente. O segundo critério comparou a técnica com as outras baseadas em decomposição de matrizes, LSI e NMF.

Para a realização dos testes foram selecionados quatro diferentes conjuntos de documentos[2]. O primeiro conjunto tem 108 artigos abordando dez conceitos da área de inteligência artificial. O segundo, 139 artigos relacionados a *Clustering*, *Biclustering*, *Ontology*, *Latent Semantic Indexing*, *Information Retrieval*, *Ontology Extraction*, *Fuzzy* e *Semantic Web*. O terceiro conjunto tem artigos apenas sobre Algoritmos Genéticos; o último conjunto tem artigos coletados aleatoriamente, sem restrição de área do conhecimento.

[2]Disponíveis em `http://www.dca.fee.unicamp.br/~ricarte/ACT`

Os resumos desses artigos foram submetidos a um estágio de preparação, que selecionou os termos mais relevantes e gerou as matrizes D e M. Também durante a preparação são definidos os valores adequados para os parâmetros do algoritmo.

A Figura 5 apresenta a interface gráfica principal do software de extração de relações semânticas e ilustra como o usuário pode supervisionar o processo. Inicialmente, o usuário ajusta os parâmetros do algoritmo como indicado em 1. Os parâmetros são de dois tipos: existem os parâmetros para o algoritmo de minimização que é baseado em técnicas evolutivas e estão indicados em 2, e existem os parâmetros da função objetivo que estão indicados em 3. Ajustados os parâmetros, o usuário inicia a busca por um protótipo de conceito por meio do botão *Play* indicado em 4. O protótipo de conceito é apresentado no painel indicado por 5, permitindo o usuário avaliar a relevância desse protótipo. Neste caso, o usuário também pode ser auxiliado pela informação de contexto do painel 9 que mostra a taxa de sobreposição com os conceitos do contexto. Se o conceito for aprovado, o usuário pode utilizar o campo 6 para aplicar um nome e armazená-lo no contexto com o botão 7. Ainda, o usuário pode alterar o conteúdo de um protótipo por meio do botão 8 que possibilita a inserção ou remoção de termos ao protótipo.

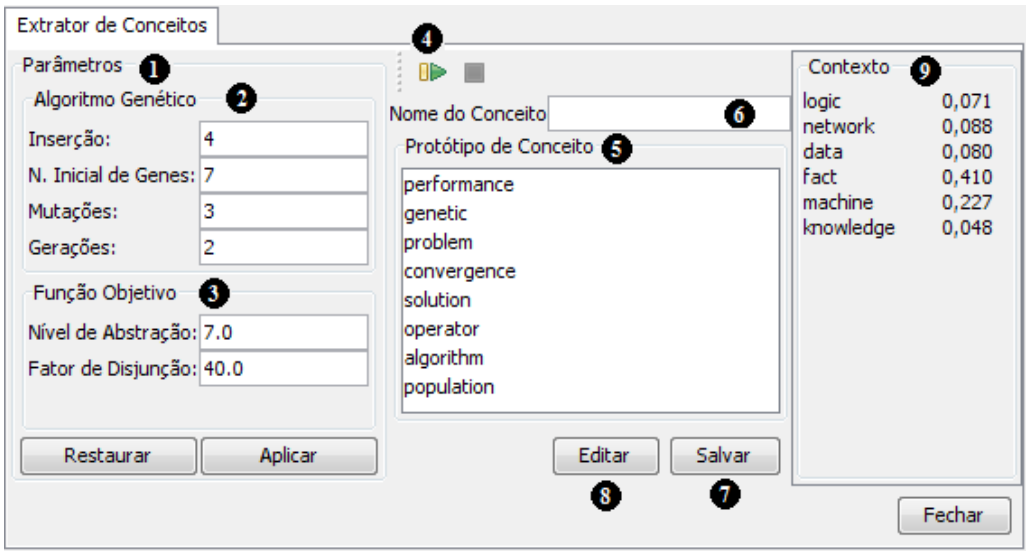

Figura 5. Interface de extração de conceitos

A proposta ACT é baseada em técnicas de agrupamento na qual a determinação automática do número de grupos é reconhecidamente um problema em aberto [Akkaya et al., 2006], que depende dos parâmetros de entrada do algoritmo de agrupamento e de fatores subjetivos, tais como: o formato e o tamanho dos grupos. Assim, a avaliação do número adequado de conceitos é difícil e não é passível de validação visual, pois os termos não podem ser representados em um plano euclideano bidimensional. Portanto, será aceitável um número que esteja próximo da quantidade real de conceitos tratados pela base de documentos. Para o primeiro e segundo conjuntos de documentos, o algoritmo identificou sete conceitos. Para o terceiro, o algoritmo identificou corretamente a existência de apenas um conceito e, para o último conjunto, o algoritmo identificou 12 conceitos. A técnica mostrou-se capaz de identificar agrupamentos coerentes de termos, reunindo termos comuns no domínio de cada conceito. Por exemplo, para o primeiro conjunto de documentos, o algoritmo identificou os agrupamentos {*neural*, *network*, *su-*

pervised, *unsupervised*}, {*genetic*, *algorithm*, *population*, *operator*} e {*data*, *mining*, *topic*, *database*}, que representam os conceitos *Neural Networks*, *Genetic Algorithm* e *Data Mining*, respectivamente.

A comparação com as duas técnicas de decomposição de matrizes foi realizada utilizando o primeiro conjunto de documentos e, para esse, a tabela 1 mostra a semelhança entre os conceitos extraídos. As linhas dessa matriz representam os conceitos obtidos com a técnica ACT e as colunas, os conceitos obtidos com NMF e LSI. O conteúdo de cada célula da matriz é um índice de semelhança entre os conceitos, cujo valor está compreendido entre 0, nenhuma semelhança, e 1, total semelhança. A semelhança entre os conceitos é definida pelo cosseno do ângulo formado entre os vetores conceitos, considerando que o conceito pode ser expresso na forma vetorial em que cada coordenada representa um termo e seu valor representa a associação fuzzy. Para facilitar a comparação entre as técnicas, células que possuem os maiores valores de semelhança em cada linha ou em cada coluna foram destacadas; a ocorrência de linhas e colunas com mais de uma marcação indica que há sobreposição entre conceitos. Quando nenhuma sobreposição é constatada, caracteriza-se um mapeamento unívoco entre os conceitos obtidos pelas duas técnicas.

	NMF								LSI						
	1	2	3	4	5	6	7		1	2	3	4	5	6	7
1	0,67	0,42	0,38	0,56	0,46	0,47	0,52		0,55	0,49	0,53	0,52	0,47	0,44	0,52
2	0,63	0,96	0,41	0,41	0,49	0,42	0,37		0,51	0,55	0,53	0,46	0,48	0,49	0,49
3	0,40	0,42	0,99	0,41	0,50	0,50	0,40		0,47	0,48	0,53	0,48	0,50	0,52	0,46
4	0,42	0,42	0,40	0,99	0,51	0,47	0,41		0,43	0,55	0,53	0,72	0,53	0,46	0,44
5	0,44	0,44	0,47	0,52	0,98	0,43	0,44		0,48	0,52	0,55	0,71	0,59	0,45	0,41
6	0,48	0,39	0,48	0,50	0,45	0,98	0,41		0,46	0,58	0,53	0,54	0,48	0,58	0,55
7	0,40	0,40	0,42	0,40	0,45	0,43	0,96		0,40	0,46	0,46	0,55	0,54	0,52	0,45

Tabela 1. Comparação entre as técnicas ACT, NMF e LSI

A comparação entre as técnicas mostra que, em relação à informação extraída, as técnicas ACT e NMF são semelhantes. A diagonal formada pelas células demarcadas mostra que existe um mapeamento um a um entre os conceitos obtidos pelas duas técnicas. O mesmo não é observado quando se compara ACT e LSI. Nesse caso há sobreposição entre conceitos, como pode ser visto analisando-se as colunas 2, 3, 4 e as linhas 5, 6.

Pela análise desses resultados, verifica-se que a técnica ACT é capaz de extrair as mesmas informações que as modernas técnicas de decomposição de matrizes com vantagens, tais como a possibilidade de supervisão do processo e definição de métricas para avaliar o número adequado de conceitos. Assim como as demais técnicas estatísticas, a proposta ACT faz uso de métricas de proximidade entre termos para a identificação de conceitos. Por essa razão, situações na qual os conceitos estão muito relacionados podem dificultar o processo de identificação dos mesmos.

5. Conclusão

Este trabalho apresentou uma nova técnica para a identificação de relações semânticas a partir da mineração de um conjunto de textos, denominada ACT — análise de correlação de termos. ACT mostrou-se bastante promissora, pois permitiu um agrupamento coerente de termos em conceitos, similar ao obtido pela técnica NMF, que nesse aspecto é superior à técnica LSI. Adicionalmente, ACT permite a intervenção de um especialista do domínio

a cada iteração, seja pela inspeção do conceito extraído pelo algoritmo de agrupamento, seja pela introdução de novos conceitos manualmente.

A implementação aqui apresentada utilizou documentos em inglês, sendo os termos identificados com o apoio da base WordNet. Embora essa base não seja adequada para identificar termos de um domínio específico, tal deficiência é sanada pela introdução de termos por uma lista complementar. Cabe destacar que essa dependência em relação à língua é apenas léxica, pois não há qualquer processamento sintático envolvido na identificação de termos e conceitos. Assim, para alterar o idioma, basta substituir o módulo WordNet por outra base lexical e modificar as listas associadas.

O algoritmo de agrupamento utiliza uma função objetivo que incorpora múltiplos aspectos, ponderados por parâmetros que devem ser ajustados pelo especialista para a obtenção de resultados coerentes. O impacto desses parâmetros na qualidade dos resultados ainda precisa ser melhor estudado. Entretanto, deve-se observar que outros tipos de função objetivo podem ser utilizados em conjunto com esse algoritmo, abrindo assim outras oportunidades de extensão deste trabalho. Outro ponto que requer maior investigação diz respeito a validação dos resultados por um especialista no domínio dos documentos.

Referências

K. Akkaya, C. Tunc, D. Aktas, and A. Altintas. On the number of clusters in channel model. In *Proceedings of the Spread Spectrum Techniques and Applications*, pages 6–9, 2006. doi: 10.1109/ISSSTA.2006.311723.

Y. Cheng and G. Church. Biclustering of expression data. *In Proc. ISMB´00*, pages 93–103, 2000.

S. T. Dumais, M. W. Berry, and G. W. O. Brien. Using linear algebra for intelligent information retrieval. *SIAM*, pages 573–595, 1995.

B. Fortuna, M. Grobelnik, and D. Mladenic. *System for semi-automatic ontology construction*, volume 4289, pages 121–131. Springer Berlin / Heidelberg, 2006.

M. Gonzalez and V. L. S. Lima. Recuperação de informação e processamento de linguagem natural. *XXIII Congresso da Sociedade Brasileira de Computação*, 3:347–395, 2003.

M. A. Hearst. *Automated Discovery of WordNet Relations,in WordNet: an electronic lexical database*. MIT Press, 1998.

Y.-J. Horng, S.-M. Chen, Y.-C. Chang, and C.-H. Lee. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *IEEE T. Fuzzy Systems*, 13(2):216–228, 2005.

D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42:378–383, 1991.

R. Pereira, I. Ricarte, and F. Gomide. Information retrieval with FROM: The fuzzy relational ontological model. *International Journal of Intelligent Systems*, 24:340–356, 2009.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513–523, 1988.

F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons. Document clustering using nonnegative matrix factorization. *Journal on Information Processing and Management*, pages 373–386, 2006.

P. Velardi, M. Missikoff, and R. Basili. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118220.1118225.