Abordagem não supervisionada para Extração de Conceitos a partir de Textos

Silvia Maria Wanderley Moraes, Vera Lúcia Strube de Lima
PUCRS - Programa de Pós-Graduação em Ciência da Computação
Av. Ipiranga, 6681, Prédio 32
Porto Alegre, Brasil 90619-900
+55 51 3320-3611

{silvia.moraes, vera.strube}@pucrs.br

ABSTRACT

This paper presents an investigation about concepts extraction from texts using clustering algorithms. We applied a hybrid approach to select feature candidates and the CLUTO tool to support the process of clustering of terms. The analysis of identified concepts was manual. The details and preliminaries results of this approach for portuguese texts are discussed.

RESUMO

Este artigo apresenta um estudo sobre extração de conceitos a partir de textos usando algoritmos de agrupamento. Utilizamos uma abordagem hibrida para selecionar os termos candidatos e a ferramenta CLUTO para apoiar o processo de clusterização de termos. A análise dos conceitos identificados foi feita manualmente. O artigo apresenta o detalhamento desse processo e discute resultados preliminares em textos da língua portuguesa.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning – *Concept learning*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*.

General Terms

Experimentation.

Keywords

Concept learning, Conceptual clustering, Feature selection, Text mining.

1. INTRODUÇÃO

Estruturas conceituais como dicionários, tesauros, taxonomias e ontologias têm se tornado um importante recurso para sistemas de informação. Em sistemas de recuperação de informações, por exemplo, tais estruturas têm ajudado a minimizar problemas de vocabulário e a aumentar a relevância das informações recuperadas. As estruturas provêem um suporte semântico que permite modificar a consulta do usuário, substituindo termos[1] desconhecidos do sistema por sinônimos, bem como enriquecendo-a com novos termos do domínio relacionados.

Em tarefas de organização da informação, como classificação e agrupamento de documentos, as estruturas conceituais têm trazido ganhos de precisão. Elas têm sido usadas para melhorar tais tarefas enriquecendo os textos com novos conceitos do domínio [3]; auxiliando na determinação da classe ou grupo através da desambiguação dos termos[1] dos documentos [2]; reduzindo a dimensionalidade de representação dos textos, através da substituição de conjuntos de termos pelos conceitos que os representam; e, ainda, ressaltando as características mais adequadas a partir das relações semânticas entre os termos [5].

Embora seja grande a aplicabilidade das estruturas conceituais é alto o custo de sua construção e manutenção. Por esta razão, e devido ao grande volume e riqueza de documentos textuais digitais disponíveis atualmente, pesquisas têm sido realizadas com o objetivo de construir tais estruturas automaticamente. As abordagens propostas geralmente usam informações lingüísticas e estatísticas sobre os termos no texto, para extrair os conceitos e as suas relações semânticas. É comum também o uso de algoritmos de aprendizagem de máquina, principalmente de agrupamento, para identificar os conceitos a partir da co-ocorrência de termos em um contexto, e para construir as estruturas conceituais propriamente ditas, usando abordagens hierárquicas [7].

A construção de estruturas conceituais tem como ponto de partida a identificação e extração dos conceitos presentes nos textos, e é este o objetivo do nosso trabalho. Baseamos nossa abordagem de extração de conceitos em trabalhos bem conhecidos como o da Hipótese de Distribuição de Harris (na qual os termos são considerados similares se compartilham os mesmos contextos) e também nos trabalhos que caracterizam os contextos a partir de relações morfossintáticas como os de Grefenstette [9] e de Hindle [10].

Utilizamos as dependências sintáticas entre os verbos e seus argumentos para identificar os termos e multi-termos (n-gramas) relevantes nos textos. A seleção dos termos é realizada através de uma abordagem híbrida que combina as medidas *Tfidf* [13] e *C-Value* [6]. Os conceitos são obtidos e organizados através de algoritmos de agrupamento disponíveis na ferramenta CLUTO [11].

[1] Entendemos "termo" como uma unidade atômica de significado, que pode ser representado por uma palavra, o radical de uma palavra, um sintagma nominal, etc.

Nossos experimentos foram realizados em 4.407 documentos da seção esportes do *corpus* PLN-BR CATEG[2], e a qualidade semântica dos *clusters* foi avaliada manualmente.

Este documento está organizado em 4 seções. Na Seção 2 discutimos dois trabalhos relacionados ao nosso. Na Seção 3 apresentamos algumas características do *corpus* e a metodologia usada em nossos experimentos. Na Seção 4 apresentamos nossas considerações finais.

2. TRABALHOS RELACIONADOS

Os trabalhos nessa área são recentes e, portanto, ainda em desenvolvimento. Yang e Callan em [17] identificam conceitos em uma coleção de mensagens eletrônicas, e utilizam técnicas de mineração de textos, recuperação de informações, processamento de linguagem natural e aprendizagem de máquina para gerar uma ontologia de conceitos. Em sua abordagem usam n-gramas e a máquina de busca do *Google* para identificar os conceitos mais relevantes. São considerados importantes aqueles termos que ocorrem com maior freqüência (acima de um determinado limiar). As relações de hiperonímia são identificadas a partir da *WordNet* e um algoritmo de agrupamento supervisionado baseado no *k-medoids* é usado para gerar a hierarquia de conceitos. Os resultados obtidos pelos autores são motivadores.

Sung e demais autores em [15] apresentam um algoritmo de agrupamento baseado no ciclo de vida de eventos na *web*. Certos eventos geram publicações na *web*, tais como novos artigos. Tipicamente, no início essas publicações são pequenas, crescem e gradualmente acabam desaparecendo. De acordo com os autores, um novo relacionamento entre conceitos pode ser criado toda vez que um certo evento acontece. Para extrair os conceitos, os autores utilizam técnicas de processamento de linguagem natural e recuperação de informações, incluindo *parsers*, algoritmo de *stemming*, remoção de *stopwords* e identificação de multi-termos. Os autores propõem um algoritmo de agrupamento conceitual baseado nos eventos da *web*. De acordo com esses autores, a abordagem pode ser usada também para enriquecer ontologias.

Os trabalhos mencionados utilizam abordagens semelhantes para extração de conceitos, ambas com o uso de n-gramas e empregando algoritmos de aprendizagem de máquina.

3. METODOLOGIA

Assim como em [15] e [17], usamos uma abordagem extensional para extrair os conceitos, e consideramos cada termo como um conceito candidato que pode ser formado por múltiplas palavras. Nas subseções seguintes, nossa metodologia é detalhada.

3.1 *Corpus* PLN-BR CATEG

Em nossos experimentos foram usados 4.407 documentos da seção esportes do *corpus* PLN-BR CATEG. O *corpus* PLN-BR CATEG possui cerca de 9.780.220 *tokens* e é composto por 30 mil textos em Língua Portuguesa do jornal Folha de São Paulo dos anos de 1994 a 2005. Escolhemos o domínio esportes, pois de acordo com [1] e [12], o vocabulário dos textos dessa seção é mais constante e reduzido se comparado às demais seções jornalísticas do *corpus*. Julgamos essas características também adequadas para a construção de uma estrutura conceitual. Os textos de esportes possuem 1.635.623 de *tokens* com uma média de 371 *tokens* por texto e com 26.143 termos diferentes (*types*).

3.2 Seleção dos Termos

Inicialmente, os textos da seção esportes foram pré-processados pela ferramenta FORMA[3]. A ferramenta segmenta o texto, lematiza e atribui etiquetas morfológicas para palavras e sinais de pontuação, com precisão em torno de 95% [8]. Optamos pela lematização como forma de normalização lingüística porque, para idiomas com morfologia flexional mais complexa como o Português, algoritmos simples de *stemming* costumam não ser suficientes, além de terem um custo computacional mais alto [16].

Em seguida, através de um *shallow parser* que desenvolvemos, são identificados os sintagmas nominais (SNs) que desempenham papel de sujeito, objeto direto ou indireto. Foram considerados aqueles SNs que aparecem em ao menos 2 documentos. Visando reduzir o número de termos, usamos duas medidas, *Tfidf* e *C-Value* para atribuir pesos aos termos e selecioná-los.

A escolha por uma abordagem híbrida foi baseada em trabalhos recentes [4][18] que apresentam resultados animadores ao combinarem mais de uma medida de ponderação dos termos. Escolhemos *Tfidf* por ser uma medida bem conhecida, usada nos trabalhos citados. Já a medida *C-Value* foi escolhida por parecer adequada para extrair multi-termos. Ela combina tanto conhecimento lingüístico quanto estatístico ao definir a importância dos termos [14].

A medida *C-Value* estima o peso de um termo *t* usando características como a sua freqüência absoluta *f(t)* no *corpus* e o seu comprimento |*t*| (número de constituintes). Usa, ainda, dois tipos de freqüências relacionadas ao conjunto *S(t)* que é formado por todos os termos do *corpus* que contenham um termo "aninhado" (*nested*) de *t*. Os termos aninhados de *t* são subconjuntos de *t*. Por exemplo, o termo "campeão do torneio de tênis" possui os termos aninhados: "campeão do torneio", "campeão de tênis", "torneio de tênis", "campeão", "torneio" e "tênis". Enquanto a freqüência *f(s)* contabiliza a ocorrência de termos aninhados de *t* em outros termos candidatos, a |*S(t)*| apenas determina a quantidade de termos candidatos que contêm um termo aninhado de *t*. A Equação 1 apresenta o método de cálculo da *C-Value* [6].

$$C-Value(t) = \begin{cases} log_2(|t|) \times f(t) \;\; se \;\; S(t) = \emptyset \\ log_2(|t|) \times f(t) - \frac{1}{|S(t)|} \sum_{s \in S(t)} f(s) \;\; se \;\; S(t) \neq \emptyset \end{cases} \quad (1)$$

Em textos jornalísticos é muito comum a referência a datas, períodos e dias da semana, tornando esses termos muito freqüentes. Por esta razão, eliminamos esses termos. Foram desprezados também artigos, pronomes, numerais e caracteres especiais. Dentre as preposições, a única que foi considerada foi "de", por ser a mais freqüente no corpus (~50%) e fazer parte de multi-termos.

Determinados os pesos, a exemplo de Butters e Ciravegna em [4], optamos por não usar um limiar (peso mínimo) previamente estabelecido para selecionar os termos. Em geral, limiares fixos não levam em consideração a real distribuição dos termos.

[2] O *corpus* PLN-BR CATEG foi constituído a partir do projeto "Recursos e Ferramentas para Recuperação de Informações em Bases Textuais em Português do Brasil" (PLN-BR). Mais informações sobre o projeto podem ser encontradas em http://www.nilc.icmc.usp.br/plnbr/.

[3] www.inf.pucrs.br/~linatural/ferramentas.htm

Utilizamos um limiar baseado na soma da média aritmética μ dos pesos p(t) dos *n* termos do *corpus* (Equação 2) com o seu desvio padrão σ (Equação 3).

$$\mu = \frac{\sum p(t)}{n} \quad (2)$$

$$\sigma = \sqrt{\frac{\sum p(t) - \mu}{n}} \quad (3)$$

Para reduzir ainda mais a quantidade termos, podemos aumentar o limiar selecionando termos cuja variação do peso em relação à média é ainda maior. Com esse propósito Butters e Ciravegna definem um *fator* que potencializa o desvio padrão (Equação 4).

$$limiar = fator \times \sigma + \mu \quad (4)$$

Para definir este *fator*, analisamos a similaridade entre os termos do *corpus*, par a par, usando a medida co-seno como métrica de similaridade. Optamos por usar *fator = 2* pois, conforme a Figura 1, é com esse fator que conseguimos maior similaridade entre os termos.

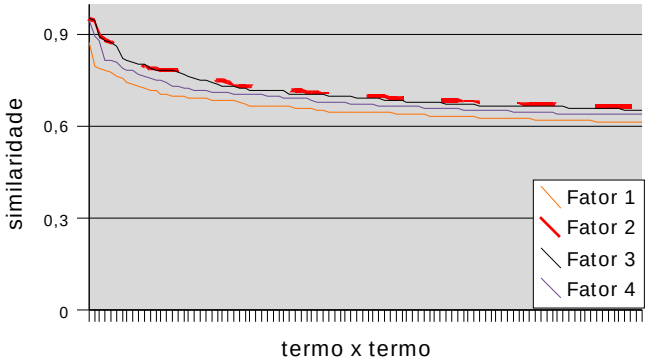

Figura 1. Fatores e similaridade entre os termos.

Calculamos, então, dois limiares: um para os pesos *Tfidf* e outro para os pesos *C-Value*. Com esses limiares realizamos os cortes, selecionando *TF* termos a partir da medida *Tfidf* e *TC* termos a partir da medida *C-Value*. Os termos selecionados são os que pertencem a $TF \cup TC$. Dos 22.691 termos reconhecidos da seção esportes foram selecionados 278 pela *Tfidf* e 109 pela *C-Value*, sendo que 88% dos termos selecionados pela *Tfidf* eram unigramas e que 37% e 60% dos termos selecionados pela *C-Value* eram, respectivamente, bigramas e trigramas. As medidas selecionaram 31 termos em comum, totalizando 356 termos a serem agrupados para formar conceitos.

3.3 Extração dos Conceitos

Para descobrirmos os conceitos presentes nos textos de esportes, usamos a ferramenta CLUTO [11] para agrupar os termos selecionados. CLUTO[4] é um pacote de software que contém um conjunto de algoritmos de agrupamento, bem como ferramentas para analisar e visualizar os *clusters* encontrados. Para viabilizar o uso da ferramenta, construímos uma matriz 356 x 356 de similaridades, onde as linhas correspondem aos termos a serem agrupados e as colunas expressam as similaridades entre os termos.

Usamos a medida co-seno para calcular essas similaridades, analisando os contextos dos termos. Os contextos são representados como *bag-of-words* cujos pesos são as frequências de co-ocorrência. Para determinar a configuração mais adequada, analisamos de forma empírica os resultados de diferentes configurações e optamos pelo uso do método de agrupamento de partição *rbr*, que realiza sucessivas bissecções, subdividindo cada grupo gerado em outros dois. Como métrica de similaridade usamos a medida *corr* que define o coeficiente de correlação entre os objetos a serem agrupados.

A ferramenta CLUTO exige também que seja informada a quantidade *k* de grupos desejada. Em função disso, realizamos testes para *k = {10, 20, 30, 40, 50, 60, 70}*. Para cada teste, analisamos manualmente a qualidade dos *clusters*, determinando: a quantidade de *clusters* que caracterizavam um conceito, a quantidade de *clusters* indefinidos (aqueles sem uma semântica clara entre os termos ou com mais 45% de termos não relacionados) e o número médio de termos semanticamente dissociados dos demais termos do *cluster*. A Figura 2 apresenta os resultados dessa análise.

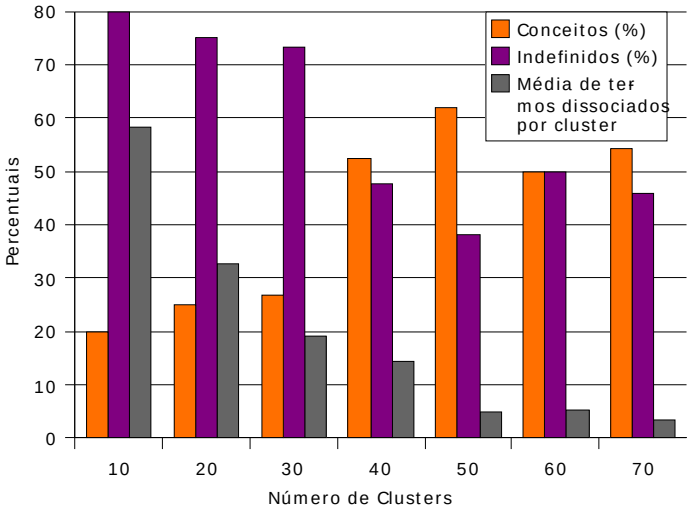

Figura 2. Análise dos *clusters*.

Com *k=50*, conseguimos o melhor índice de conceitos bem construídos e proporcionalmente menos *clusters* desconhecidos. A Tabela 1 exibe 10 dos 38 núcleos de conceitos encontrados com *k=50*, a maioria dos quais expressam conceitos, demonstrando a viablidade da metodologia.

No entanto, a metodologia precisa ser aperfeiçoada, em razão da inclusão de termos dissociados: "ala pivô" no *cluster* 4, "norte" no *cluster* 8, "alemão" e "inglês" no *cluster* 7. O primeiro caso é um erro e se deve a uma falha do *parser* que reconhece n-gramas. Falhas desse tipo são esperadas e, neste experimento ficaram em torno de 2%. Os demais casos são conseqüência do tipo de *corpus* que estamos usando. Em textos jornalísticos é comum citar as regiões em que os jogos acontecem, bem como comentar a nacionalidade dos pilotos mais conhecidos, tornando tais termos muito freqüentes.

[4] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

O *cluster* 2 apresenta outros problemas. Ele refere-se claramente ao domínio do futebol, no entanto nem todos os termos são significativos, como por exemplo "direita" e "esquerda". No caso de "direita" e "esquerda", o problema foi não consideramos outras preposições além do "de". Os termos "direita" e "esquerda" geralmente indicam os lados do campo e é comum associá-los a "ataque" e "jogada" gerando termos como "ataque pela direita", "jogada pela esquerda".

Os *clusters* 5, 9 e 10 apresentam similaridade léxica. Nos *clusters* 5 e 9, por exemplo, todos os termos iniciam com "vice" e no *cluster* 10, com "lateral". Isso é conseqüência do uso da medida *C-Value*, que trabalha com variações de um mesmo termo (termos aninhados) e atribui a eles o mesmo contexto. Essa é uma característica interessante da medida *C-Value* mas pode conduzir a problemas semânticos, induzindo agrupamentos incorretos. Em alguns de nossos experimentos com $k \neq 50$, observamos, por exemplo, que todos os termos dos *clusters* 5 e 9 foram colocados em um único agrupamento.

Tabela 1. Alguns conceitos encontrados com *k=50*

#	Conceito
1	estadio , ingresso, local, pessoa, publico, seguranca, torcedor
2	area, ataque, bola, cabeca, defesa, direita, entrada, esquerda, gol, goleiro, jogada, lance, linha, marcacao, penalti, placar, saida de bola
3	arbitro, cartao, cartao amarelo, falta, juiz
4	ala, ala pivo, basquete, esquema de jogo, pivo
5	vice, vice de futebol, vice presidente, vice presidente de futebol, vice presidente de marketing
6	contusao, coxa direita, joelho direito, joelho esquerdo
7	alemao, carro, categoria, circuito, prova, companheiro de equipe, corrida, ingles, piloto, piloto de teste, pista, pole position
8	norte , ranking, tenis, tenista, tie break
9	vice campea, vice campeao, vice campeonato, vice lider, vice lideranca
10	lateral, lateral direito, lateral esquerdo

4. CONSIDERAÇÕES FINAIS

Neste artigo apresentamos uma metodologia para extrair conceitos de textos automaticamente. A abordagem híbrida usando *TfIdf* e *C-value*, bem como o método de seleção de termos mostraram-se adequados na escolha de termos relevantes para a identificação de conceitos. Apesar de relatarmos vários problemas na geração dos grupos de conceitos, os resultados mostram a viabilidade da metodologia.

Acreditamos que o uso de um *corpus* de textos com definições conceituais poderá trazer resultados ainda melhores. Além de experimentos com outros *corpora*, julgamos necessário introduzir alguma metodologia para estimar o número de *clusters*, bem como métricas para a avaliação dos conceitos gerados. Também faz parte de nossos trabalhos futuros a construção de uma estrutura conceitual relacionando os conceitos identificados nos textos.

5. REFERÊNCIAS

[1] Azeredo, S., Moraes, S.M.W., and Strube de Lima, V.L. Keywords, k-NN and Neural Networks: a Support for Hierarchical Categorization of Texts in Brazilian Portuguese. In: 6th International Language Resources and Evaluation (LREC'08), Marrakech, may 28-30. European Language Resources Association (ELRA), Morocco, 2008.

[2] Bang, S. L, Yang, J.D and Yang, H. J. Hierarchical Document Categorization with k-NN and concept-based thesauri. Information Processing and Management, Nº 42, Elsevier, 2006, pp. 387-406.

[3] Bloehdorn, S., Cimiano P. and Hotho, A. Learning Ontologies to Improve Text clustering and Classification. In: 29th Annual Conference of the German Classification Society (GfKl 2005): From Data and Information Analysis to Knowledge Engineering, Magdeburg, Germany, March 9-11, 2005. Studies in Classification, Data Analysis, and Knowledge Organization, 30, Springer, pp. 334-341, February 2006.

[4] Butters, J. and Ciravegna, F. Using Similarity Metrics for Terminology Recognition. In; 6th International Language Resources and Evaluation (LREC'08), Marrakech, may 28-30. European Language Resources Association (ELRA), Morocco, 2008.

[5] Edmonds, A. Using conceptual structures for efficient document comparison and location. In: IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), Honolulu, April 1-5. IEEE Symposium Series on Computational Intelligence 2007, Hawaii, USA, 2007, pp 238-242.

[6] Frantzi, K. T. and Ananiadou, S. The C/NC value domain independent method for multi-word term extraction. Journal of Natural Language Processing 6, 3, 1999, 145–179.

[7] Gamallo, P., Lopes, G.P. and Agustini, A. Inducing Classes of Terms from Text. In: 10th International Conference Text, Speech and Dialogue (TSD 2007), Pilsen, Czech Republic, September 3-7. Lecture Notes in Computer Science, 4649, Springer, 2007, pp. 31-38.

[8] Gonzalez, M.A.I and Strube de Lima, V.L. Tools for Normalization: An Alternative for Lexical Normalization. In: International Conference on Computational Processing of Portuguese, E. Vieira et. al (eds): PROPOR 2006, Lecture Notes in Computer Science, 3960, Springer-Verlag, 2007, pp. 100-109.

[9] Grefenstette, G. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In Branimir Boguraev and James Pustejovsky (eds), Corpus processing for Lexical Acquisition, MIT Press, USA, 1996, pp. 205-216.

[10] Hindle, D. Noun classification from predicate-argument structures. In: 28th Annual Meeting of the Association of Computational Linguistics, ACL, Pittsburgh, Pennsylvania, USA, 1990, pp. 268-275.

[11] Karypis, G. CLUTO: A clustering Toolkit. University of Minnesota, Department of Computer Science, Minneapolis, Technical Report 02-017. Available from http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/manual.pdf (2003), accessed june 2008.

[12] Moraes, S.M.W. e Strube de Lima, V.L. Um Estudo sobre Categorização Hierárquica de uma Grande Coleção de Textos em Língua Portuguesa. In: V Workshop em Tecnologia da Informação e Linguagem Humana, XXVII Congresso da SBC, 5-6 julho, SBC, Rio de Janeiro, 2007.

[13] Salton, G. Introduction to Modern Information Retrieval. New York: McGraw- Hill, 1983.

[14] Spasic, I., Nenadic, G. and Ananiadou, S. Using Domain-Specific Verbs for Term Classification. In: Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan, ACL, 2003, pp. 17-24.

[15] Sung, S., Chung, S. and McLeod,D. Efficient Concept clustering for Ontology Learning using an Event Life Cycle on the Web. In: ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, March 16-20, 2008, pp. 2310-2314.

[16] Vilares, J., Barcala, F.M. and Alonso, M.A. Using Syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In: Internacional Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2002, Mexico, February 17-23. Lectures Notes in Computer Science, 2276, Springer-Verlag, 2002, pp. 381-390.

[17] Yang, H. and Callan, J. Ontology Generation for Large Email Collections. In: 9th Annual International Conference on Digital Government Research, Partnerships for Public Innovation, DG.O 2008, Montreal, Canada, May 18-21. ACM International Conference Proceeding Series, 289, Digital Government Research Center, 2008, pp. 254-261.

[18] Zhang, Z., Iria, J., Brewster, C. and Ciravegna, F. A Comparative Evaluation of Term Recognition Algorithms, In: 6th International Language Resources and Evaluation (LREC'08), Marrakech, may 28-30. European Language Resources Association (ELRA), Morocco, 2008.