Expansão de consulta por pseudo realimentação no Modelo TR+ para Recuperação de Informação

Thyago Bohrer Borges
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, prédio 32
sala 507
Porto Alegre, Brasil
+55 51 3320-3621
thyago.borges@gmail.com

Marco Gonzalez
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, prédio 32
sala 505
Porto Alegre, Brasil
+55 51 3320-3558
marco.gonzalez@inf.pucrs.br

Vera Lúcia Strube Lima
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, prédio 32
sala 507
Porto Alegre, Brasil
+55 51 3320-3621
vera.lima@inf.pucrs.br

ABSTRACT
This work presents the specification of experiments that apply query expansion techniques with pseudo relevance feedback to TR+ Model in information retrieval. The TR+ Model uses terms and binary lexical relations (BLRs) for indexing and searching of texts in Portuguese. The experiments add (or remove) new terms or BLRs to the original query in order to study the effects of those changes in document retrieval.

RESUMO
Este trabalho apresenta a especificação de experimentos para aplicação da técnica de expansão de consulta com pseudo realimentação de relevantes ao Modelo TR+ em recuperação de informação. O Modelo TR+ além de termos também se utiliza de relações lexicais binárias (RLBs) para indexação e busca de textos em língua portuguesa. Os experimentos adicionam (ou retiram) novos termos ou RLBs à consulta original com o objetivo de analisar os efeitos dessas alterações na recuperação dos documentos.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Query Formulation; Relevance Feedback – *pseudo relevance feedback.*

General Terms
Pseudo Relevance Feedback, Relevance Feedback, Expansion Query, Information Retrieval, Binary Lexical Relations, TR+ Model, Natural Language Processing.

Keywords
Recuperação de Informação, Relações Lexicais Binárias, Expansão de Consultas, Realimentação de Relevantes, Modelo TR+.

1. INTRODUÇÃO
Sistemas de Recuperação de Informação (RI), que trabalham com documentos textuais, possuem como principal objetivo atender a consultas realizadas por usuários através de indexação, busca e classificação de documentos [1]. Duas direções de pesquisa podem ser consideradas: (i) quanto à formulação da consulta e (ii) quanto à representação dos conceitos presentes nos documentos.

A maior dificuldade enfrentada pelo usuário, quanto à formulação adequada da consulta, é a decisão de quais palavras-chave possibilitam encontrar os documentos que necessita. Uma formulação eficiente passa pelo conhecimento do usuário sobre o domínio do tema a ser recuperado e sobre o próprio funcionamento do sistema. Entretanto, formular uma consulta eficiente através de palavras-chave, que possibilitem retornar informações relevantes, pode não ser uma tarefa fácil. Segundo Baeza-Yates [1], a identificação da real necessidade do usuário é um processo muito complexo e pode ser a diferença entre uma recuperação eficiente e uma recuperação que não atende as suas necessidades. Uma alternativa é a utilização de Expansão de Consulta (EC). A EC reformula a consulta original para melhorar seu desempenho.

Quanto à representação dos conceitos presentes nos documentos, diversas alternativas têm sido desenvolvidas e algumas incluem técnicas de Processamento da Língua Natural (PLN). Neste sentido, Gonzalez [3] apresentou um modelo para recuperação de informação denominado TR+. O Modelo TR+ alia métodos estatísticos a conhecimento lingüístico para indexar e recuperar textos em língua portuguesa. Ele utiliza termos e relações lexicais binárias como descritores de conceitos.

Com o objetivo principal de especificar experimentos para aplicar EC ao Modelo TR+, este artigo apresenta na seção 2 as características do Modelo TR+; na seção 3, a técnica de EC com pseudo realimentação de relevantes aplicada ao Modelo TR+; na seção 4, a especificação dos experimentos realizados; e na seção 5, as próximas etapas do trabalho a serem realizadas.

2. MODELO TR+
O Modelo TR+ [3] apresenta características próprias de um sistema de recuperação de informações textuais. É proposta a utilização de tratamento idêntico para os textos dos documentos e para as consultas formuladas em linguagem natural. Estas últimas são reformuladas com a inclusão de operadores booleanos antes do processo de busca e classificação.

O Modelo TR+ indexa e recupera documentos utilizando, tanto na formulação da consulta quanto na indexação, descritores de conceitos que incluem termos simples e compostos [6]. O primeiro passo é o pré-processamento do texto, onde são utilizados métodos de tokenização e etiquetagem morfológica. Após é realizada a nominalização e a captura das relações lexicais binárias (RLBs),

Nominalização, no Modelo TR+, é o processo de transformação de adjetivos, verbos e advérbios em substantivos [5]. Com ela definem-se os termos simples que constituirão os descritores. As RLBs [4] constituem os termos compostos e completam a

descrição dos conceitos presentes nos documentos. RLBs (dos tipos classificação, restrição e associação) são relacionamentos entre termos nominalizados, que capturam mecanismos de coesão frásica [3].

Os descritores (termos e RLBs) têm seus pesos calculados através do conceito de evidência. Neste cálculo é considerada a freqüência de ocorrência dos termos e, também, o número de relações que há entre eles.

3. PSEUDO REALIMENTAÇÃO DE RELEVANTES

EC é uma técnica utilizada para aperfeiçoar os resultados obtidos por consultas realizadas por usuários em um sistema de RI. Esta técnica agrega à consulta original novos termos que, em princípio, devem ser relevantes para as informações que o usuário pretende encontrar. Uma técnica muito utilizada de EC é a Realimentação de Relevantes (*Relevance Feedback*) [7]. A idéia principal é utilizar o usuário para avaliar o resultado inicial de sua consulta, julgando os documentos recuperados e indicando quais são relevantes. Essas informações, que podem ser termos ou expressões [8], são agregadas à consulta original buscando obter melhores resultados.

Uma evolução da técnica realimentação de relevantes utiliza a realimentação das informações sem a participação do usuário. Esta técnica é denominada Pseudo Realimentação de Relevantes [9]. Nela, após o processo de recuperação inicial, assume-se que os *n* documentos melhores classificados são relevantes e deles são extraídos termos ou expressões que realimentarão de forma automática a consulta inicial [7].

4. EXPERIMENTOS

Para a realização dos experimentos foi utilizado como ponto de partida os resultados obtidos por Gonzalez [3] através do Modelo TR+ sem EC. Foi empregada a metodologia utilizada nas TRECs (*Text Retrieval Conferences*), sendo utilizada como coleção de documentos 4156 artigos do Jornal Folha de São Paulo de 1994. Foram realizadas 50 consultas referentes a 50 tópicos distintos, sendo cada um deles representado por um título, uma descrição e uma narrativa (que consiste de características que identificam um documento relevante). Após a realização das consultas, foi utilizado o método de *pooling* [2] para julgar a relevância dos documentos recuperados. A avaliação dos resultados obtidos é finalizada com o cálculo de métricas consolidadas para se avaliar sistemas de RI: precisão, abrangência, média harmônica entre precisão e abrangência, além da medida MAP [2].

De posse das consultas e dos documentos recuperados para cada consulta respectivamente, realizamos até o presente momento 7 experimentos incluindo EC ao Modelo TR+. Na Figura 1 apresentamos o processo dos experimentos realizados.

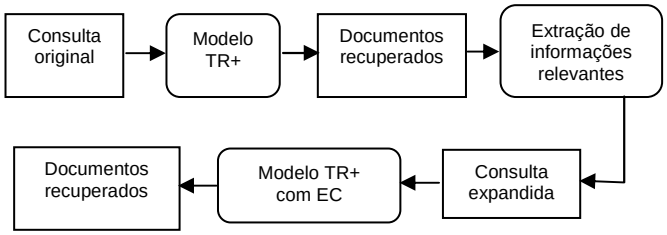

Figura 1. Processo realizado para os experimentos.

Todos os experimentos consideraram os três documentos do topo da classificação, em relação à consulta original, para realizar a extração das informações relevante (Figura 1). O experimento A1 inclui na EC as três RLBs com maior peso. O experimento A2 utilizou as três RLBs com maior peso de acordo com seu tipo (restrição, associação e classificação). O experimento A3 utilizou os três termos com maior peso.

O experimento A4 utilizou uma abordagem diferente. Foram retiradas as RLBs da consulta original para verificar a influência que elas têm na recuperação.

O experimento A5 considerou as cinco RLBs com maior peso. O experimento A6 considerou os cincos termos com maior peso. E finalmente o experimento A7 considerou as dez RLBs com maior peso.

4.1 Resultados e considerações parciais

Nesta seção apresentaremos os resultados obtidos até o momento pelos experimentos com EC junto ao Modelo TR+ e os resultados obtidos pelo Modelo TR+ sem a EC, com o objetivo de compararmos tais resultados (Tabela 1). Os resultados são apresentados utilizando a precisão (Pr) dos documentos recuperados para cada intervalo da abrangência (Abr) dos mesmos no intervalo entre 0 (0%) e 1 (100%).

Podemos exemplificar melhor utilizando os valores apresentados na Tabela 1 para o experimento A: quando a abrangência está na faixa de valores entre 0 (0%) e 0,1 (10%) a sua precisão é de 0,9733 (%).

Tabela 1. Resultados dos experimentos realizados

	A	A1	A2	A3	A4	A5	A6	A7
Abr	Pr	Pr	Pr	Pr	Pr	Pr	Pr	Pr
0	0,973	0,973	0,973	0,297	0,965	0,793	0,227	0,973
0,1	0,973	0,972	0,972	0,269	0,955	0,792	0,187	0,972
0,2	0,962	0,956	0,956	0,248	0,925	0,776	0,186	0,956
0,3	0,955	0,940	0,940	0,247	0,897	0,760	0,186	0,940
0,4	0,929	0,894	0,894	0,239	0,854	0,719	0,179	0,894
0,5	0,924	0,874	0,874	0,237	0,841	0,704	0,177	0,874
0,6	0,883	0,804	0,804	0,229	0,773	0,655	0,175	0,804
0,7	0,835	0,765	0,765	0,20	0,743	0,619	0,168	0,765
0,8	0,771	0,706	0,706	0,177	0,684	0,565	0,159	0,706
0,9	0,607	0,546	0,546	0,085	0,542	0,433	0,068	0,546
1	0,483	0,392	0,392	0,051	0,392	0,32	0,039	0,392
	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP
	0,850	0,808	0,808	0,169	0,777	0,652	0,113	0,808

Os resultados obtidos até então parecem indicar que a EC não traria benefícios ao Modelo TR+ (a medida MAP fica em média, 4 pontos percentuais menor com EC nos melhores casos). Uma explicação para tal comportamento é o fato de que RLBs são características particulares de certos documentos (aproximadamente 90% das RLBs ocorrem em apenas um documento) e, uma vez que RLBs relevantes, oriundas dos documentos recuperados, são incluídas na consulta original, estas RLBs só fortalecem a recuperação dos mesmos documentos, sem melhorar a recuperação. Por outro lado, RLBs não relevantes incluídas, obviamente, prejudicam a recuperação.

5. PRÓXIMAS ETAPAS DO TRABALHO

Este trabalho apresentou a especificação de experimentos para a aplicação de EC com pseudo realimentação de relevantes ao Modelo TR+.

Planejamos a expansão das consultas utilizando os usuários para definir quais documentos são realmente relevantes a cada consulta original. Documentos estes de onde serão extraídas as informações (ternos ou RLBs) que serão adicionadas às consultas originais, justificando assim a utilização da técnica Realimentação de Relevantes [7]. Os experimentos que realizaremos com a presença dos usuários para definição dos documentos relevantes terão uma etapa a mais do que os experimentos apresentados neste trabalho. A etapa será o julgamento por parte dos usuários de quais são os documentos que deverão ser extraídas as informações que expandiram as consultas originais. Futuramente poderemos confrontar os resultados obtidos pelas tuas técnicas de EC e os resultados alcançados por Gonzalez [3].

6. AGRADECIMENTOS

Estudo realizado pelo Centro de Desenvolvimento e Pesquisa PUCRS-DELL. Termo Aditivo: Programa de Pesquisa e Desenvolvimento em Tecnologia da Informação - PDTI 001/2008, financiado pela Dell Computadores do Brasil Ltda. com recursos da Lei 8.248/91.

7. REFERÊNCIAS

[1] Baeza-Yates, R. A. and Ribeiro-Neto, B. 1999 Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc.

[2] Buckley, C. and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM, New York, NY, p.25-32.

[3] Gonzalez, M. (2005). Termos e relacionamentos em evidência na recuperação de informações. Tese de Doutorado, Universidade Federal do Rio Grande do Sul, Br.

[4] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. (2005). Binary Lexical Relations for Text Representation in Information Retrieval. In: International Conference on Applications of Natural Language to Information Systems. Springer Verlag, p.21-31. (Lectures Notes in Computer Science, 3513).

[5] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. (2006) Tools for Nominalization: an Alternative for Lexical Normalization. In: Workshop on Comp. Proc. of the Portuguese Lang. - Written and Spoken, 7; PROPOR, 2006. Springer-Verlag, p.100-109. (Lectures Notes in AI, 3960).

[6] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. (2006). Lexical Normalization and Relationship Alternatives for a Term Dependence Model in Information Retrieval. In: Comp. Ling. and Intel. Text Processing - Int. Conf., 7; CICLing, 2006. Springer-Verlag, p.394-405. (Lectures Notes in Computer Science, 3878).

[7] Manning C. D., Raghavan P., Schütze H. (2008). in press. Introduction to Information Retrieval. Cambridge: Cambridge University Press.

[8] White, R. W., Marchionini, G. (2006). A study of real-time query expansion effectiveness. SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (p. 715-716). New York, NY, USA: ACM Press.

[9] Youngjoong Ko, H. A., Seo, J. (2007). An efective snippet generation method using the pseudo relevance feedback technique. SIGIR '07: Proceedings of the 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (p. 711-712). New York, NY, USA: ACM Press.