Utilizando Co-Training para Realimentação de Relevância na *WEB*

Matheus Victor Brum Soares[1], Ronaldo C. Prati[2], and Maria Carolina Monard[1]

[1] Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo, Campus de São Carlos – USP/São Carlos
Av. Trabalhador São-Carlense, 400 / CEP 13566-590 / São Carlos-SP
[2] Centro de Computação, Matemática e Cognição - CMCC
Universidade Federal do ABC - UFABC
Rua Santa Adédia, 166 / Santo André-SP
{caneca,prati,mcmonard}@icmc.usp.br

Resumo Ao realizar buscas na *WEB*, diferentes usuários podem utilizar as mesmas palavras-chave com objetivos de busca distintos. Para que seja possível uma maior personalização da busca, os critérios de relevância do usuário devem ser levados em consideração. O processo de realimentação de relevância permite que o usuário indique, dentre os sites encontrados, quais ele considera como relevantes ou não, para que seja possível a reordenação dos resultados da busca. Essa reordenação deve colocar sites semelhantes aos considerados relevantes dentre os primeiros, e sites semelhantes aos considerados irrelevantes dentre os últimos. Para atingir esses objetivos, neste trabalho é proposto o C-Search, que é uma ferramenta que realiza essa reordenação, utilizando o algoritmo de aprendizado parcialmente supervisionado multivisão co-training. Experimentos iniciais mostram que, para consultas genéricas mas que possuam um bom diferencial entre sites relevantes e irrelevantes, o sistema consegue obter bons resultados.

Key words: Realimentação de Relevância, Aprendizado Parcialmente Supervisionado Multivisão, Mineração de Textos

1 Introdução

Cada vez mais, pessoas precisam encontrar informações em sites, livros, jornais, revistas e outras fontes armazenadas digitalmente, utilizando a *WEB* como principal fonte de busca. Todavia, recuperar a informação desejada pelo usuário é uma tarefa que apresenta várias dificuldades, pois, para obter documentos de seu interesse, o usuário deverá traduzir seu critério de busca em uma consulta apropriada utilizando, por exemplo, palavras-chave. Entretanto, em muitos casos, buscas realizadas dessa maneira não são suficientes para satisfazer as necessidades do usuário em um sistema de busca. Um problema comum nesses casos é a alta presença de documentos não relevantes entre os documentos retornados. Em

outras palavras, o conjunto de respostas a uma consulta contém a maioria dos documentos relevantes ao usuário (alta cobertura[3]), mas contém também um grande número de documentos irrelevantes (baixa precisão). Nesse cenário, os principais objetivos dos sistemas de busca são recuperar o maior número possível de documentos relevantes (alta cobertura) com o menor número possível de documentos não relevantes (alta precisão).

Uma maneira de contribuir para alcançar esses objetivos é processar os resultados de uma consulta de maneira que os documentos irrelevantes sejam filtrados. Entre as abordagens propostas para a construção desses filtros encontram-se os algoritmos de Aprendizado Parcialmente Supervisionado [1] (APS)[4]. Neste trabalho é proposto o uso de APS, mais especificamente, aqueles que utilizam visões múltiplas (multivisão[5]) do conjunto de dados, para a construção de filtros de realimentação de relevância para a recuperação de informação. Algoritmos de aprendizado parcialmente supervisionado são particularmente úteis nesse processo, pois requerem poucos exemplos rotulados. Além disso, a multivisão pode auxiliar no processo de categorização dos dados.

Neste trabalho são descritos experimentos iniciais realizados com a ferramenta C-Search, por nós desenvolvida, que realiza realimentação de relevância utilizando o algoritmo de APS multivisão co-training. A proposta foi inicialmente avaliada utilizando um pequeno número de consultas livres realizadas por um grupo de voluntários. Os resultados experimentais mostraram bons resultados para consultas pouco específicas nas quais os sites relevante e irrelevantes contêm informações suficientes que os diferenciem, considerando as poucas preferências indicadas pelo usuário.

Este trabalho está organizado da seguinte maneira: na Seção 2 a realimentação de relevância é descrita em linhas gerais; na Seção 3 é brevemente descrito o algoritmo co-training; na Seção 4 é descrita a ferramenta C-Search; na Seção 5 são apresentados alguns resultados experimentais; e na Seção 6 são apresentadas as conclusões deste trabalho.

2 Realimentação de Relevância

Somente o usuário pode dizer se os resultados de uma dada consulta lhe são relevantes ou não. Dessa maneira, para aumentar a efetividade da recuperação de informação, foi proposto um processo de Realimentação de Relevância (RR) [2]. Ele consiste em o usuário identificar (rotular), de acordo com suas necessidades, alguns documentos relevantes e irrelevantes dentre o conjunto de documentos retornados na consulta. Sistemas de RR funcionam de forma iterativa. Inicialmente, um sistema de busca recupera um conjunto de documentos referentes a uma certa consulta. O usuário seleciona alguns poucos documentos relevantes

[3] Cobertura, do inglês *recall*

[4] Também denominado de aprendizado semissupervisionado

[5] Multivisão dos dados, ou multidescrição, é motivada pelo fato que, frequentemente, existem conjuntos de informações diferentes que se referem ao mesmo objeto, e cada conjunto de informação é capaz de descrever de maneira independente o objeto.

e irrelevantes entre os retornados, e o sistema de RR pode construir uma nova consulta baseada nestes documentos, ou reordenar os documentos já retornados. Nesse processo, o sistema de RR deveria recuperar mais documentos semelhantes aos documentos relevantes, e/ou ordenar documentos mais relevantes antes de documentos menos relevantes.

Existem inúmeras técnicas de realimentação de relevância. Neste trabalho o foco está nos métodos que utilizam aprendizado de máquina. Tendo alguns poucos documentos rotulados como relevantes ou irrelevantes, o processo de RR pode se enquadrar no cenário de aprendizado parcialmente supervisionado, considerando o problema de realimentação de relevância como um problema de classificação com poucos exemplos rotulados e um número expressivo de exemplos não rotulados. O sistema C-Search proposto utiliza esses exemplos rotulados para reordenar o conjunto de documentos, de tal maneira que os documentos similares àqueles que o usuário indica como relevantes possam ser mostrados primeiro a ele. Para rotular esses exemplos o C-Search utiliza o algoritmo co-training, resumidamente descrito a seguir.

3 Co-Training

co-training [3] é um algoritmo de aprendizado parcialmente supervisionado multivisão que consiste na indução de dois ou mais classificadores, sendo que cada um deles é induzido utilizando uma descrição diferente dos exemplos de treinamento. Esses classificadores coopeream entre si para rotular exemplos cujos rótulo (classe) não são conhecidos. Essa cooperação entre os classificadores pode ser implementada de diferentes maneiras. Neste trabalho, dois classificadores são gerados para duas descrições diferentes do conjunto de exemplos, e um exemplo do conjunto não rotulado é rotulado apenas se houver um alto grau de confiança na classificação desse exemplo por esses dois classificadores.

Esses novos exemplos automaticamente rotulados são então adicionados ao conjunto original de exemplos rotulados, e o processo é repetido até rotular todos os exemplos ou, como no caso da cooperação entre os classificadores implementada neste trabalho, até que não seja possível rotular exemplos com alto grau de confiança. Nesse último caso, após a execução de co-training, tem-se um conjunto maior de exemplos rotulados, o qual pode ser utilizado como conjunto de treinamento por qualquer outro algoritmo de aprendizado supervisionado para induzir um novo classificador.

4 C-Search

Para avaliar a hipótese de que algoritmos de aprendizado parcialmente supervisionado multivisão podem ser usados no processo de realimentação de relevância, foi implementada a ferramenta C-Search [4] para reordenação dos documentos resultantes de uma busca. Essa ferramenta reorganiza os resultados de uma busca na *WEB* visando personalizar o resultado final a partir dos interesses do usuário, utilizando APS multivisão.

Na Figura 1 é mostrado uma visão geral da ferramenta. A partir de uma busca realizada pelo usuário (I), o motor de busca[6] (II) retorna os sites que contêm os resultados iniciais dessa busca (III). Esses três passos são realizados comumente no dia a dia da maioria dos usuários da Internet ao realizarem buscas em um sistema de busca qualquer. Analizando esses resultados iniciais, o usuário deve rotular alguns poucos exemplos (sites) como relevantes ou irrelevantes utilizando somente sua opinião pessoal quanto à relevância do assunto que está buscando (IV). Com essa informação, o sistema C-Search gera duas visões V_1 e V_2, baseadas no título e na descrição do site (V). Essas duas visões são processadas pela ferramenta de pré-processamento de textos PreTexT II[7], que utiliza a abordagem *bag-of-words* (VI) para a criação das duas tabelas atributo-valor correspondentes (VII). Essas tabelas são então utilizadas pelo co-training, no qual o *Naïve Bayes* foi utilizado como algoritmo base ($VIII$) para reordenar os sites da busca inicial, obtendo então o resultado final (IX), em que esses sites são mostrados ao usuário na ordem de relevância encontrada pelo C-Search. Os passos IV à IX representam a realimentação de relevância realizada pela ferramenta C-Search.

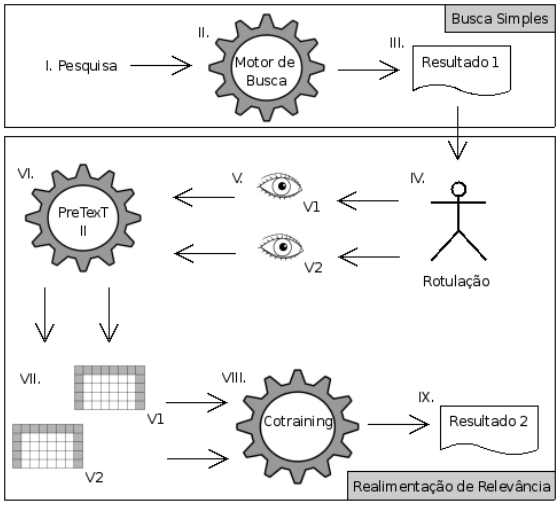


Figura 1. Visão geral da ferramenta C-Search

[6] Neste trabalho foi utilizado o *Yahoo! Search API* como motor de busca.
[7] `http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_333.pdf`

5 Experimentos

Para avaliar a nossa proposta, foi solicitado a um grupo de usuários voluntários realizar uma consulta **genérica** no buscador C-Search[8]. Cada busca realizada pelo usuário retorna 100 (cem) *WEB* sites. A partir dos sites retornados, o usuário deve indicar alguns poucos sites como relevantes ($\oplus$) ou irrelevantes ($\ominus$), conforme sua opinião pessoal. Nos experimentos realizados, foi solicitado que os usuários rotulassem 4 (2 $\oplus$ e 2 $\ominus$), 5 (3 $\oplus$ e 2 $\ominus$, ou 2 $\oplus$ e 3 $\ominus$) e 6 (3 $\oplus$ e 3 $\ominus$) sites. Para fins de avaliação, o usuário deve ao final indicar dentre todos os 100 sites retornados quais são relevantes e quais são irrelevantes para que seja possível construir as curvas recall-precisão para caracterizar o desempenho do C-Search para aquela consulta.

Os gráficos recall-precisão agrupam os resultados em grupos de dez sites. Cada gráfico contêm três curvas: **Normal** que representa o resultado de recall-precisão do motor de busca apenas, **Rank 1** e **Rank 2** que representa o resultado do primeiro e segundo *rankings* implementados no C-Search, respectivamente. **Rank 1** é definido como o produto do grau de confiança da classificação dos sites rotulados nas visões V_1 e V_2, enquanto que **Rank 2** é definido como a soma dos *rankings* dessas visões. Cada ponto na curva representa múltiplos de dez sites, ou seja, o primeiro ponto da curva representa o recall e a precisão dos 10 primeiros sites retornados, o segundo ponto representa o recall e a precisão dos 20 primeiros sites retornados, e assim por diante.

São utilizados os seguintes três critérios de comparação para avaliar o desempenho do C-Search:

Critério 1: leva em consideração se a curva de um *ranking* está acima da curva de outro *ranking*, ou seja, tem a precisão superior em toda, ou a maior parte da curva. Em casos em que as curvas estão próximas (tais como na Figura 2(b)), o resultado é considerado neutro.

Critério 2: leva em consideração o número de pontos existentes na curva. Um número menor de pontos indica que o *ranking* posicionou todos os sites relevantes dentre os primeiros do *ranking*, deixando somente sites irrelevantes no final do *ranking*.

Critério 3: leva em consideração somente os primeiros pontos das curvas no gráfico. Ou seja, se a precisão do primeiro ponto da curva for maior, significa que o usuário obteve mais resultados relevantes logo no primeiro conjunto de 10 sites mostrados ao usuário na tela, *i.e.* top-10. Nesse caso o usuário não necessitaria procurar por mais resultados em outras telas.

Na Tabela 1 encontram-se as 18 consultas genéricas realizadas pelos voluntários[9], o número de sites relevantes totais indicados pelos respectivos usuários dentre os 100 sites retornados pela busca, assim como os sites relevantes dentre os 10 primeiros sites, e os sites relevantes dentre os 20 primeiros sites retornados antes da realimentação de relevância, *i.e.* utilizando o *Yahoo! Search API* como motor de busca.

[8] `http://sistemas.labic.icmc.usp.br:8088/realimentar/`

[9] As consultas são apresentadas examente como digitadas pelos usuários.

Usuário	Consulta	Relevantes		
		Totais	Primeiros 10	Primeiros 20
01	caneca	17	3	4
02	cotinga	26	2	6
03	Direito Processual do Trabalho	27	3	8
04	maquinas finita	9	**7**	7
05	proficiencia ingles	24	5	8
06	sarmiento	17	3	4
07	aprender chinês	29	**7**	11
08	mercado de derivativos	93	**10**	19
09	Nero	9	2	2
10	Álcman	38	**6**	13
11	Complexo MHC	46	**7**	13
12	esporte	8	3	5
13	fender	14	**6**	7
14	freesbe	38	**7**	10
15	machine learning evaluation	16	4	7
16	segunda guerra mundial	27	5	9
17	Thomas Mann	30	**7**	12
18	títulos palmeiras	21	**8**	14

Tabela 1. Consultas e número total e parcial de sites relevantes marcados pelo usuário por consulta.

Após executar a realimentação de relevância no C-Search e analizando os gráficos recall-precisão de cada consulta[10] foi possível classificar os resultados das consultas 1 a 6 como positivos[11], de 7 a 9 como neutros e de 10 a 18 como negativos.

Para ilustrar, na Figura 2(a) são mostrados os resultados correspondentes à consulta `sarmiento`[12], com a indicação do usuário de 3 sites relevantes e 3 sites irrelevantes. O C-Search obteve 90% de precisão logo no primeiro grupo de sites, o que poderia deixar o usuário satisfeito com a consulta. Assim essa consulta foi classificada como positiva — Tabela 1.

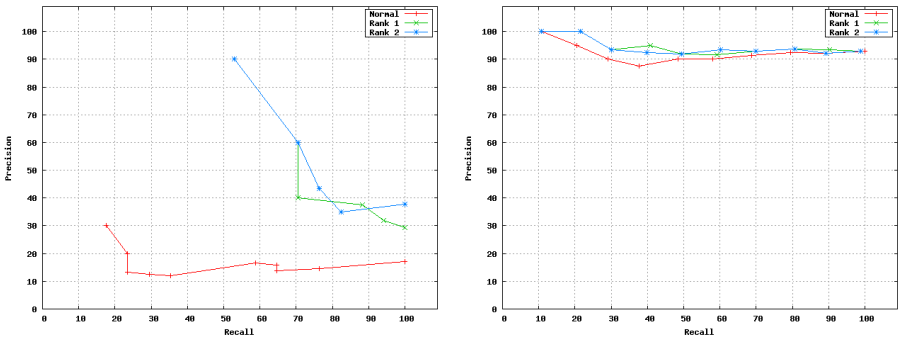
Na Figura 2(b) é mostrado o resultado neutro obtido pela consulta `mercado de derivativos`, na qual o C-Search obteve resultados similares ao motor de busca. Observe que para esta consulta, quase todos os sites (93 na Tabela 1) foram considerados relevantes pelo usuário, podendo ser considerado um caso típico no qual a RR é desnecessária.

Quanto a resultados negativos, na Figura 3(a) encontra-se o gráfico recall-precisão da consulta `segunda guerra mundial`, para o qual o usuário indicou três sites relevantes e dois sites irrelevantes. Observando os sites indicados pelo usuário como relevantes e irrelevantes, foi possível notar que ambos continham informação, ainda que parcial, relacionada com a consulta do usuário, *i.e.* os sites não são "fortemente relevantes/irrelevantes". Assim, o C-Search foi novamente executado mas indicando sites fortemente relevantes/irrelevantes no mesmo número que anteriormente. Os resultados dessa segunda consulta são

[10] Os gráficos recall-precisão de todos os testes realizados pelos usuários podem ser encontrados em `http://labic.icmc.usp.br/resultados/caneca`.

[11] *i.e.* o C-Search obteve resultados melhores dos obtidos pelo motor de busca utilizado.

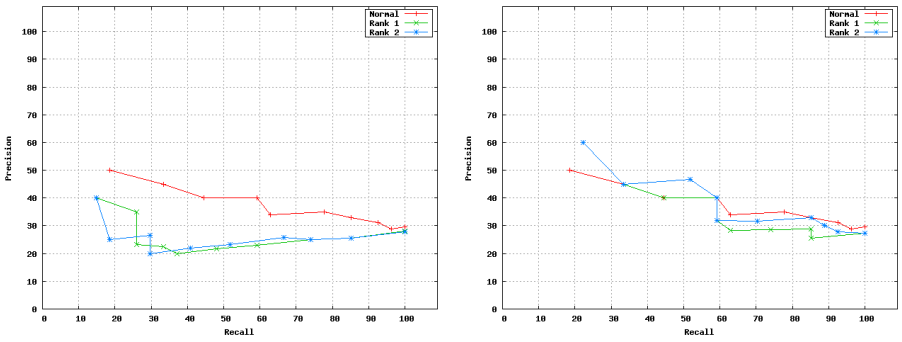
[12] Domingo Faustino Sarmiento foi presidente da Republica Argentina (1968-1974), destacando-se pela sua atuação na área de educação



(a) (b)

Figura 2. Gráficos recall-precisão das consultas (a) `sarmiento`, (b) `mercado de derivativos`

mostrados na Figura 3(b) os quais, ainda que poderiam ser classificadas como neutros, mostram que para o **Rank 2**, o primeiro conjunto de dez sites contêm mais sites relevantes que os retornados inicialmente pelo motor de busca, e aproximadamente 50% dos resultados relevantes para este usuário podem ser encontrado entre os 30 primeiros sites.



(a) (b)

Figura 3. Gráficos recall-precisão da consulta `segunda guerra mundial` (a) realizada pelo usuário, (b) realizada utilizando melhores exemplos

Deve ser observado que na avaliação experimental realizada com 18 usuários, a RR foi utilizada independentemente dos resultados obtidos pelo motor de busca. Nesse caso, dos 18 testes realizados 6 (33.3%) foram positivos, 3 (16, 7%) neutros e 9 (50.0%) negativos. Porém, em uma aplicação real, caso o resultado de uma busca simples seja satisfatório, dificilmente o usuário utilizaria uma ferramenta para melhorar a busca. Considerando que a obtenção de mais da metade de sites relevantes no primeiro grupo de dez sites retornados pelo buscador é

satisfatório (Primeiros 10 > 5 na Tabela 1 indicados em negrito) então provavelmente em somente 9 dos testes realizados o usuário ativaria a opção de RR, com os seguintes resultados: 5 (55.6%) positivos, 1 (11.1%) neutro e 3 (33.3%) negativos.

6 Conclusão

Neste trabalho foi apresentada a ferramenta C-Search, que realizada realimentação de relevância para buscas na web utilizando aprendizado parcialmente supervisionado multidescrição. Resultados experimentais iniciais mostraram que, nos casos em que o conjunto de sites relevantes e irrelevantes têm características bem definidas, o C-Search consegue obter bons resultados na reordenação das consultas. Entretanto, é importante que os sites indicados sejam "fortemente" relevantes ou "fortemente" irrelevantes. Foi observado que após indicar os sites relevantes, o usuário é menos cuidadoso na indicação dos sites a serem rotulados como irrelevantes. Em outras palavras, não há a preocupação em distinguir sites pouco relevantes de sites muito (fortemente) irrelevantes. Como a qualidade dos poucos exemplos rotulados é um fator importante para o algoritmo de aprendizado parcialmente supervisionado obter bons resultados, a realimentação de relevância realizada com usuários que têm esse comportamento fica prejudicada.

Foi observado que o C-Search obteve bons resultados principalmente em pesquisas ambíguas que retornam sites fortemente irrelevantes, e nas quais o usuário não encontra muitos sites relevantes no primeiro grupo de 10 sites. Vale observar que essas são as situações nas quais há interesse em utilizar a realimentação de relevância. Acreditamos que, nesses casos, como o usuário deve analisar um número maior de sites procurando os sites relevantes para rotular, encontra mais sites fortemente irrelevantes durante a procura pelos relevantes. Dessas maneira, o problema descrito previamente é aliviado. Considerando o número de consultas analisadas (18 consultas livres) os resultados são promissores, mas são resultados iniciais. Assim, um número maior de consultas deverá ser analisada futuramente para chegar a resultados mais conclusivos.

Agradecimentos: Aos revisores do trabalho pelas sugestões e críticas. Trabalho realizado com auxílio do CNPq.

Referências

1. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. Cambridge, MA: MIT Press. (2006)
2. Ruthven, I., Lalmas, M.: A survey on the use of relevance feedback for information access systems. In: The Knowl. Eng. Rev., vol 18 (2), pp. 95–145. (2003)
3. Blum, A., Mitchell T.: Combining labeled and unlabeled data with co-training. In: 11th Annual Conference on Computational Learning Theory (COLT), pp. 92–100. ACM Press New York (1998)
4. Soares, M.V.B.: Aprendizado de máquina parcialmente supervisionado multidescrição para realimentação de relevância em recuperação de informação na *WEB*. Tese de Mestrado, ICMC-USP. (2009)