Ranqueamento Supervisionado de Autores em Redes de Colaboração Científica

Paulo J. L. Alvarenga[1], Marcos A. Gonçalves[2], Daniel R. Figueiredo[3]

[1] Universidade Federal de Itajubá
[2] Universidade Federal de Minas Gerais
[3] Universidade Federal do Rio de Janeiro
pauloalvarenga@unifei.edu.br, mgoncalv@dcc.ufmg.br, daniel@land.ufrj.br

Abstract. The problem of ranking in collaboration networks consists in determining an ordering of researchers according to their influence or prestige using network metrics. This paper proposes a supervised machine learning approach that combines four metrics to rank nodes. Experiments using a database of Brazilian researchers in Computer Science and taking as reference the Research Productivity scholarships from CNPq assigned to researchers show 11% gains on average precision in comparison with results obtained with state-of-the-art metrics used in isolation.

Resumo. O problema de ranqueamento em redes de colaboração científica consiste em definir uma ordenação dos pesquisadores de acordo com sua influência ou prestígio utilizando métricas da rede. Este artigo apresenta uma abordagem baseada em aprendizagem de máquina supervisionada que combina quatro métricas para produzir o ranqueamento. Experimentos realizados em uma base de dados com pesquisadores brasileiros da área da Ciência da Computação tendo como referência a bolsas de Produtividade em Pesquisa do CNPq alocadas aos pesquisadores, demonstraram ganhos de até 11% na precisão média em relação aos resultados obtidos com métricas do estado-da-arte quando utilizadas isoladamente.

Categories and Subject Descriptors: H. Information Systems [**H.m. Miscellaneous**]: Machine Learning

General Terms: Machine Learning, Authors Ranking, Collaboration Network

Keywords: Supervised Learning, Ranking, Collaboration Network

1. INTRODUÇÃO

Redes de colaboração científica representam alguma relação de cooperação entre um conjunto de pesquisadores ou grupos de pesquisa. Uma destas redes mais estudadas é formada por vértices que representam indivíduos (pesquisadores) e arestas que representam publicações científicas em co-autoria. Ou seja, dois indivíduos estão relacionados (possuem uma aresta) se eles são co-autores de ao menos um artigo científico. Arestas destas redes geralmente possuem pesos que refletem a intensidade da colaboração entre os indivíduos. O número de artigos publicados em conjunto normalizado pelo número de co-autores de cada artigo é um exemplo de peso para capturar a intensidade da colaboração [Newman 2004a].

Diferente finalidades levam ao estudo de redes de colaboração científica, tais como indentificação de grupos coesos, recomendação para colaborações, e ranqueamento de indivíduos ou grupos. Neste último, o objetivo é definir um ranqueamento que capture a infuência ou prestígio dos pesquisadores utilizando principalmente métricas obtidas da rede de colaboração científica, tais como o grau dos vértices, o peso em arestas incidentes, e distâncias entre vértices. De fato, este é um problema que vem sendo explorado na literatura [Freire and Figueiredo 2011; Newman 2004a; 2004b]. Um ranqueamento consistente de pesquisadores possui diversas aplicações, tais como ajudar nas decisões ligadas a processos de promoção, alocação de recursos financeiros, e ranqueamento de artigos científicos.

O ranqueamento de pesquisadores em redes de colaboração é geralmente realizado utilizando uma única métrica obtida da rede de colaboração [Freire and Figueiredo 2011; Newman 2004a]. Entretanto, métricas distintas quando comparadas diretamente normalmente levam a diferentes ranqueamentos [Freire and Figueiredo 2011], o que é natural uma vez que diferences métricas podem capturar diferentes aspectos da influência dos pesquisadores. Desta forma, é interesssante considerar algoritmos de ranqueamento que considerem diferentes métricas obtidas da rede de colaboração científica e as combine de acordo com aspectos estruturais do problema. De fato, esta abordagem melhor reflete o processo real de ranqueamento de pesquisadores, que geralmente combina diferentes critérios para estabelecer um ranqueamento.

Neste trabalho propomos o uso de algoritmos de *learning to rank* (aprender a ranquear) para combinar diferentes métricas de forma automática para produzir melhores ranqueamentos de pesquisadores. Algoritmos de *learning to rank* são algoritmos de aprendizado de máquina supervisionado especialmente projetados para otimizar uma determinada métrica de ranqueamento (e.g., precisão média). Em particular, iremos aplicar um algoritmo tradicional e eficiente conhecido por RankSVM [Joachims 2002; Tsochantaridis et al. 2006]. Para avaliação e comparação do método iremos considerar a base de dados de pesquisadores brasileiros da área de Ciência da Computação e as seguintes métricas:: (i) número de publicações com ao menos um co-autor; (ii) número de colaboradores; (iii) intensidade de colaboração fora do conjunto de brasileiros. Esta última métrica foi proposta em [Freire and Figueiredo 2011] e uma comparação entre as métricas (isoladamente) foi realizada no mesmo trabalho tendo como critério a alocação de bolsas de Produtividade de Pesquisa realizadas pelo CNPq. Nossos experimentos apresentam ganhos de até 11% na precisão média quando comparados com o ranqueamento obtido com as mesmas métricas consideradas isoladamente.

Esse artigo está organizado da seguinte forma. Seção 2 cobre os trabalhos relacionados, incluindo as métricas de ranqueamento definidas que seram consideradas. Seção 3 apresenta o funcionamento do algoritmo de *SVM-Rank*. Seção 4 apresenta a avaliação experimental, incluindo as métricas para avaliação do ranqueamento e discussão dos resultados obtidos. Por fim, Seção 5 apresenta conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

O problema de ranqueamento em redes de colaboração científica vem sendo discutido na literatura [Freire and Figueiredo 2011; Newman 2004a; 2004b]. Em [Freire and Figueiredo 2011] os autores propõe uma nova métrica para ranqueamento baseada na intensidade de colaboração dos pesquisadores com o exterior de um determinado conjunto (ex. conjunto dos pesquisadores brasileiros dentro da rede de todos os pesquisadores). O artigo também realiza uma comparação entre três métricas distintas utilizando a base de dados do DBLP[1] tendo como referência pesquisadores com bolsa de Produtividade em Pesquisa do CNPq. A métrica proposta se mostra superior às outras quando comparando a precisão e abrangência na recuperação de pesquisadores 1A e 1B dentro do universo de pesquisadores brasileiros.

Neste trabalho iremos combinar todas as métricas apresentadas e comparadas em [Freire and Figueiredo 2011] utilizando o algoritmo rankSVM (detalhes na seção 3) para produzir o ranqueamento dos pesquisadores. Nossa avaliação irá utilizar a mesma base de dados utilizada em [Freire and Figueiredo 2011] assim como o mesmo padrão de referência. Uma descrição sucinta das métricas e da base se encontra na Seção 4 e maiores detalhes estão disponíveis em [Freire and Figueiredo 2011]. Por fim, iremos utilizar uma outra métrica para avaliação do ranqueamento dos pesquisadores, conhecida como nDCG (descrita sucintamente na Seção 4).

[1] Disponível publicamente em http://www.informatik.uni-trier.de/~ley/db/

3. O MÉTODO DE *LEARNING TO RANK*

Nessa Seção apresentamos o funcionamento básico dos Support Vector Machines (SVMs), um método de classificação do estado-da-arte, e da extensão proposta para esse dar suporte ao ranqueamento.

3.1 Funcionamento do SVM

O SVM (*Support Vector Machine*), também conhecido como *kernel machine*, é um método de definição de um hiperplano de margem máxima (*Maximum Margin hyperplan*), que possibilita que o modelo seja escrito como a soma de influências de um subconjunto das instâncias de treino. Essas influências são dadas por núcleos de similaridade especificados pela aplicação, podendo ser lineares, radiais, entre outras [Alpaydin 2010]. O objetivo do SVM é a classificação de elementos, ou seja, identificar a classe dos elementos analisados, de acordo com a posição que o elemento se encontra no hiperplano. Como exemplo, para nosso caso de pesquisadores, poderíamos tentar descobrir se, de acordo com as entradas, um pesquisador pertence à classe 2 ou não. Sendo assim, o SVM com núcleo linear traça um hiperplano n-dimensional linear (onde n é o número de características informadas para cada indivíduo) para separar os pesquisadores de classe 2 dos demais pesquisadores. Dada a função do hiperplano para identificação da classe de cada autor, basta fornecer seus dados, e a função do SVM é posicionar o autor como antes do hiperplano – abaixo da margem negativa – como não pertencendo à classe, ou após o hiperplano – acima da margem positiva – na qual o elemento é caracterizado como pertencendo à classe analisada. A tarefa de identificar o hiperplano de corte pode ser formulada como um problema de otimização convexo, no qual existe um único ótimo que pode ser resolvido analiticamente [Alpaydin 2010].

Em alguns casos, ditos não linearmente separáveis, o hiperplano linear traz muitos erros. Para esses casos, é comum usar outros tipos de funções, como funções radiais, que é o caso deste trabalho, com o núcleo RBF (*radius based function*). A ideia é similar à do linear, porém agora o objetivo é circunscrever os elementos de determinada classe, mantendo os elementos que não pertencem à classe fora da circunscrição formulada. Neste trabalho foi usado o núcleo RBF, pois um teste preliminar usando o SVM linear (ou seja, sem uso de *kernel*) não obteve resultados satisfatórios.

3.2 Funcionamento do SVMRank

Em [Joachims 2002] é proposta uma alteração ao SVM capaz de utilizá-lo para o ranqueamento de documentos, e não somente para a classificação. A proposta é dada da seguinte forma: ao analisar o SVM, é possível transformar o mapeamento no qual um elemento é alocado a uma classe em um vetor peso P, ortogonal ao hiperplano que separa as classes. Sendo assim, o ranqueamento dos elementos pode ser dado pela distância entre o mapeamento do elemento no plano e a margem que separa a classe. Desse modo, o peso P positivo favorece o elemento quanto maior for o peso que indica que ele pertence à classe, e negativo quanto maior a distância que indica que ele não pertence à classe (menor peso). Logo, o ranqueamento é dado diretamente pelo cálculo de P, o que permite seu uso nos diversos núcleos disponíveis para o SVM, como o RBF, que foi o escolhido para este trabalho.

4. AVALIAÇÃO EXPERIMENTAL

Descrevemos inicialmente as características da base de dados utilizada, seguido pela descrição das métricas de avaliação, do procedimento experimental adotado, e finalmente dos resultados obtidos.

4.1 Descrição dos dados utilizados

A base utilizada neste artigo é a mesma de [Freire and Figueiredo 2011], onde construiu-se uma rede de colaboração científica utilizando a base de dados do DBLP obtida em julho de 2009. A DBLP é uma

base de dados pública com informações bibliográficas de periódicos e conferências principalmente da área de Ciência da Computação, e contava na época com mais de 1,3 milhões de publicações e 750 mil pesquisadores. Apesar de ser uma referência mundial e usada pela comunidade acadêmica para busca de informação bibliográfica, sua cobertura é limitada quando considerando conferências e periódicos brasileiros e algumas subáreas da Ciência da Computação. Um total de 2733 pesquisadores ligados ao Brasil foram identificados nesta rede (detalhes em [Freire and Figueiredo 2011]), entretanto, iremos considerar neste trabalho apenas 378 pesquisadores, que são os que recebiam bolsa de Produtividade em Pesquisa do CNPq em 2009.

O CNPq divide seus pesquisadores associados em duas categorias, 1 e 2, sendo que a categoria 1 é subdividida em quatro níveis, A, B, C e D. As diferentes categorias são usadas para refletir senioridade, produtividade e impacto dos pesquisadores e é também relacionado ao valor monetário do associado. A categoria 2 serve em sua maioria jovens pesquisadores, enquanto a categoria 1 requer pelo menos oito anos desde a obtenção do título de doutorado. A categoria 1A é a mais prestigiosa, e é reservada para pesquisadores que mostraram excelência continuada em produção científica e treinamento de recursos humanos, e são membros de grupos de pesquisa consolidados. A lista dos bolsistas associados é disponibilizada publicamente e mantida pelo CNPq. Esta categorização será usada como identificador da importância de cada pesquisador a ser ranqueado, logo, será considerada a métrica alvo para o ranqueamento.

4.2 Métricas de avaliação

Existem diferentes métricas que são comumente usadas para avaliar a qualidade de um ranqueamento. Neste trabalho, iremos utilizar o nDCG (*Normalized Discounted Cumulative Gain* ou ganho acumulado descontado normalizado), o qual é capaz de lidar com múltiplos níveis de relevância dos elementos do ranqueamento. Para obter-se o nDCG, primeiro obtemos o CG (ganho acumulado), distribuindo a cada elemento a ser recuperado um valor de ganho (o que chamaremos vetor G). O ganho acumulado em uma posição de ranqueamento i é computado pela soma de G nas posições 1 a i. Portanto, o CG pode ser obtido recursivamente como um vetor CG[i], definido como:

$$CG[i] = \begin{cases} G[1], & se\ i = 1; \\ CG[i-1] + G[i], & caso\ contrrio. \end{cases} \quad (1)$$

Para reduzir a pontuação do ranqueamento à medida que ele progride, porém com menores ganhos, é usada uma função de desconto, frequentemente usada na literatura como o log na base b da posição do ranqueamento, também recursivamente definida como:

$$DCG[i] = \begin{cases} CG[i], & se\ i\ <\ b \\ DCG[i-1] + G[i]/^{b}log(i) & se\ i >= b. \end{cases} \quad (2)$$

Por fim, para deixar os valores relativos à medida ideal, calcula-se o nDCG de cada posição i dividindo-se o DCG[i] pelo DCG[i] ideal (ou seja, o DCG[i] que teria precisão 100% em todos os níveis de evocação, obtido a partir do ranqueamento a partir dos ganhos em ordem decrescente). Maiores detalhes e comparações entre o nDCG e outras medidas de avaliação podem ser obtidas em [Järvelin and Kekäläinen 2002].

4.3 Procedimento Experimental

O modelo de validação utilizado em nosso experimento foi o *repeated 3-fold cross-validation*. Em conformidade com o *3-fold cross-validation*, esses dados foram divididos em três partições diferentes. Cada partição possui uma quantidade igual de autores para cada classe de bolsista CNPq[2]. Apesar de

[2]O objetivo é que cada um tivesse a mesma quantidade, nem todos os níveis de pesquisadores apresentavam quantidade múltipla de três. Em todos os casos houve resto 2, e escolhemos distribuir os restos sempre entre a primeira e a segunda

garantida a quota mínima de autores por classe de bolsista CNPq, cada partição possui um subconjunto aleatório e disjunto de autores de cada classe. Com a finalidade de garantir variedade no treino, foram criados 10 agrupamentos diferentes, cada um contendo três partições. Dessa forma é feito o *repeated 3-fold cross-validation*, no qual é repetido o processo do *3-fold cross-validation* em cada um dos 10 grupos, e ao final todo o conjunto é analisado. O processo de treinamento consiste em três fases:

(1) Geração de um modelo, na qual o modelo é treinado a partir de uma da primeira partição;
Este modelo é gerado pela ferramenta "svm_rank_learn".[3] O Kernel usado foi o RBF, que precisa ser configurado a partir de dois parâmetros: -c (*trade-off* entre o erro de treinamento e margem); e -g (parâmetro *gamma* da função radial usada na RBF). Foram feitos experimentos variando tanto c quanto gamma calculados a partir de valores 2^i, sendo c de 2^-5 a 2^15 e gamma de 2^-15 a 2^3, usando i com passo 2.

(2) Validação do modelo com a segunda partição, a fim de fazer a seleção de parâmetros, ou seja, identificar os parâmetros cuja configuração fornece o melhor resultado;
Para cada modelo gerado na etapa 1, foi executado o ranqueamento da segunda partição, e, posteriormente, calculados os valores nDCG, para obter a precisão do ranqueamento dado. O teste foi feito com o algoritmo "svm_rank_test"[Joachims 2006].

(3) Teste do modelo com a terceira partição, a fim de verificar a eficiência da configuração em um grupo de autores não analisado.
Para os melhores parâmetros identificados no passo 2, são validados os resultados na terceira partição, fazendo novamente a verificação do nDCG, para obter a confiabilidade do treinamento.

Com a finalidade de garantir a equidade dos testes, os três passos descritos foram feitos para três combinações das partições. A primeira combinação usou a primeira, segunda e terceira partições para as três etapas descritas: treino, validação e teste, respectivamente. A segunda combinação usou a segunda, terceira e primeira partições, respectivamente, e a terceira combinação, da mesma forma, usou a terceira, primeira e segunda. Ao final é contabilizada a média dos nDCGs em cada posição, que é o resultado do "*3-fold cross-validation*".

4.4 Modelagem adotada

Como iremos utilizar um algoritmo de aprendizado de máquina supervisionado, é necessário conhecer o resultado (classificação) desejado no momento do treinamento. Dessa forma, somente os 378 pesquisadores bolsistas do CNPq serão utilizados neste artigo, tanto para treinar o modelo quanto para avaliá-lo posteriormente. Além disso, a categoria do bolsista será considerado um ranqueamento natural dos pesquisadores, representando a influência ou prestígio de cada pesquisador. Ou seja, a ordenação ideal apresentaria os pesquisadores agrupados na ordem 1A, 1B, 1C, 1D, 2.

As seguintes métricas da rede de colaboração foram utilizadas pelo algoritmo (ver detalhes em [Freire and Figueiredo 2011]):

(1) número de artigos publicados (de acordo com a base da DBLP em junho de 2009);
(2) número de co-autores (grau do vértice na rede de colaboração);
(3) peso do vértice (métrica definida em [Newman 2004a]), que representa o número de publicações com ao menos um co-autor;
(4) peso do vértice no corte (métrica definida em [Freire and Figueiredo 2011]), que representa a intensidade de colaboração entre o pesquisador (brasileiro) e pesquisadores fora do conjunto dos pesquisadores brasileiros.

---

partições, deixando a terceira partição ligeiramente menor

[3]O algoritmo, assim como explicação de uso, está disponível para download em: `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

Para os testes, o conjunto de 378 pesquisadores foi dividido em três partições. As duas primeiras possuem 127 autores e a terceira partição possui 124 pesquisadores. A opção de deixar sempre a terceira partição menor tem como objetivo permitir a comparação entre todas as partições e melhor identificar se o uso de menor quantidade de autores causaria algum ruído. A seguir é mostrada a quantidade de cada classe, assim como quantos pesquisadores dessa classe foram aleatoriamente distribuídos em cada partição:

—1A: 21 autores – 7 para cada partição

—1B: 23 – 8 para as duas primeiras, 7 para a última

—1C: 32 – 11 para as duas primeiras, 10 para a última

—1D: 59 – 20 para as duas primeiras, 19 para a última

—2: 243 – 81 para cada partição

A categoria da bolsa do CNPq foi escolhida como a classe de cada pesquisador durante o treinamento. Neste raciocínio, autores 1A devem ocupar as primeiras posições do ranqueamento, seguidos por 1B, 1C, 1D, e nas últimas posições do ranqueamento estariam os pesquisadores de nível 2. Para o uso no SVM, foi distribuída a cada classe um peso decrescente, recebendo 1A, 1B, 1C, 1D e 2, respectivamente, os pesos 6, 5, 4, 3 e 2, pesos estes que foram usados no cálculo do nDCG.

Além disso, de modo a fazer com que os atributos de entrada fossem comparáveis entre si, todos foram normalizados, para assumirem um valor proporcional entre 0 e 1, de acordo com a fórmula:

$$\frac{x_j^{(i)} - min\{x_k^{(i)}, k = 1, ..., N^{(i)}\}}{max\{x_k^{(i)}, k = 1, ..., N(i)\} - min\{x_k^{(i)}, k = 1, ..., N^{(i)}\}}$$

A fórmula tem como resultado o valor normalizado de cada atributo $x_j^{(i)}$, para um dos N atributos i do autor j.

4.5 Resultados

A partir dos testes realizados foi calculado o nDCG médio de cada posição, proveniente das 10 experimentações obtidas pela validação cruzada, de acordo com os parâmetros escolhidos. Esta média foi comparada com os valores relativos ao uso das métricas geradas no experimento que usamos como base [Freire and Figueiredo 2011]. A título de comparabilidade com os resultados do artigo base, foi mantida a nomenclatura das métricas, sendo elas: *Degree*: número de colaboradores de cada autor; *w(v)* peso do vértice, de acordo com *Newman*, *q(v)* o peso de corte do vértice. O valor médio de cada nDCG calculado pela validação cruzada sobre o ranqueamento baseado em cada uma das métricas é apresentado no gráfico 1.

Ainda com base neste gráfico, analisando os mínimos e máximos de nDCG em cada posição, temos: SVM-Rank: 67% e 90%; *Degree*: 56% e 88%; *w(v)*: 61% e 89%;*q(v)*: 53% e 87%. Considerando estes números isoladamente, houve uma leve melhoria no nDCG máximo, porém o nDCG mínimo é superior ao mínimo das demais métricas, garantindo uma qualidade mínima do resultado como um todo. Além disso, foi possível obter uma combinação capaz de trazer um ranqueamento predominantemente melhor para os primeiros 15 pesquisadores (num resultado perfeito, as 15 primeiras posições teriam apenas os pesquisadores 1A seguidos pelos pesquisadores 1B). Apesar disso, a métrica w(v) ainda foi, na média, melhor que o algoritmo proposto. Isso é devido ao ranqueamento de um pesquisador nível 2 que, de acordo com todas as métricas, exceto w(v), era ranqueado em primeiro lugar. Em quase todos os experimentos este pesquisador também foi ranqueado em primeiro lugar, o que trouxe um primeiro lugar inferior ao de w(v). Apesar de indesejável, tal problema é esperado, visto que o ranqueamento aprendido é afetado diretamente pelas bases usadas para aprendizagem.

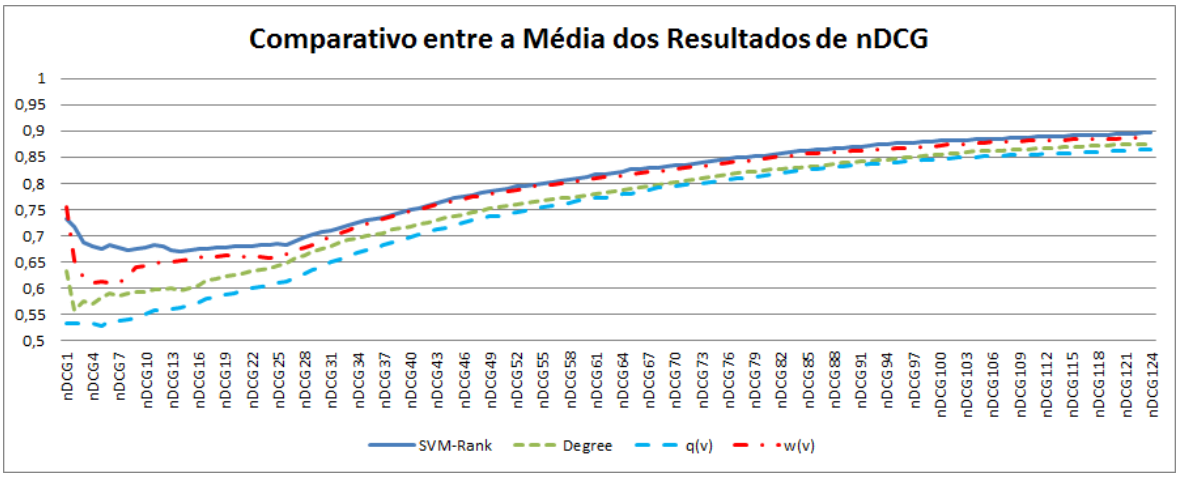

Fig. 1. Comparativo entre o método proposto e as métricas isoladas

Para validar se o uso da combinação das métricas a partir do SVM é realmente melhor, foi feita uma validação estatística usando o *Wilcoxon Signed-Rank Test*, que é um t-test para amostras correlacionadas. Foi usado o nDCG para cada uma das 124 posições, obtido pelo uso do SVM-Rank, e este foi pareado com o nDCG combinado das três métricas usadas no baseline, sendo que nesta combinação foi considerado apenas o melhor nDCG dentre os obtidos pelas métricas base para a posição específica - no nosso experimento, foi usado predominantemente o nDCG de $w(v)$. O teste pareado foi feito com todos os valores das 10 execuções, totalizando 1240 amostras de resultados de ranking. O resultado obtido foi valor $W = 426669$, $n_(s/r) = 1238$, $z = 16,96$, o que implica que o uso do SVM-Rank para ranquear baseado no uso das outras métricas com significância maior que 99,9% que apenas a escolha manual de cada um dos melhores resultados, de cada métrica considerada isoladamente. A distribuição da média desses resultados pode ser percebido no gráfico 2.

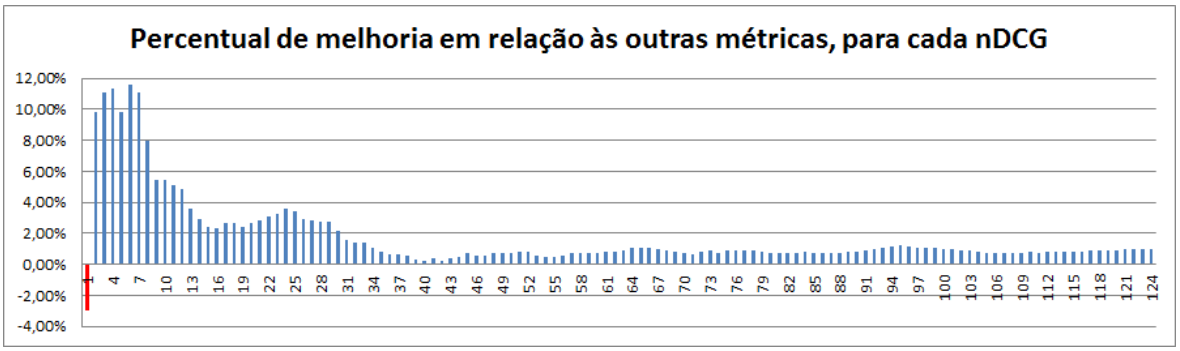

Fig. 2. Percentual de melhoria do método proposto em relação à união dos melhores resultados de todas as métricas

Neste caso, o primeiro valor do nDCG (nDCG@1) foi 2,94% pior do que as bases usadas. Mais uma vez, nota-se a influência na primeiro pesquisador recuperado ser de nível 2. Uma das possibilidades dessa ocorrência é que um pesquisador não pode ser nível 1 se ele possui menos de 8 anos de conclusão de doutorado (regra usada pela classificação CNPq), porém o perfil de publicações e colaboração em redes de pesquisa deste pesquisador seja a de um pesquisador de um nível mais alto, como 1A. Deste modo, torna-se necessário também verificar a influência do tempo de conclusão do doutorado no ranqueamento do pesquisador. No entanto, em todos os outros resultados houve melhoria, chegando

ao ápice de melhoria no nDCG@6, que se apresentou 11,55% superior à melhor entre as métricas comparadas.

5. CONCLUSÃO E TRABALHOS FUTUROS

Em nosso modelo foi possível, a partir de dados de rede de colaboração de autores, aplicar a técnicas de *learning to rank*, mais especificamente, com uso do algoritmo *SVM-Rank*, de modo a obter melhoria em ranqueamento de autores. Foi usado o método *repeated 3-fold cross validation*, a partir do qual foi possível constatar a melhoria de até 11,55% no cálculo do nDCG, comparado com os resultados das bases comparadas. Portanto, o uso do algoritmo SVM-Rank para a tarefa mostrou-se promissor.

Apesar dos resultados, é reconhecido o impacto de outros fatores na influência do ranqueamento de autores. No caso da classificação do CNPq, todos os pesquisadores com menos de 8 anos de conclusão do doutorado não podem ser classificados como nível 1. Logo, verificar o impacto da idade nessa classificação se faz necessário e é uma próxima investigação natural a ser feita.

Outro fator que influencia na classificação do CNPq é a formação de profissionais (orientações de mestrado e doutorado), que poderá ser levantada para geração do modelo.

Quanto ao algoritmo SVM-Rank, a necessidade de dados rotulados para treino limitou a quantidade de indivíduos disponível para a análise, e outro trabalho futuro é a verificação de outros métodos para obter rótulos de indivíduos automaticamente. Uma possibilidade é o uso de aprendizagem de máquina semi-supervisionada, como o uso do *Expectation-Maximization*, de modo a adicionar apenas autores rotulados artificialmente enquanto se garante o ganho de informação.

Finalmente, de modo a poder aplicar o modelo a pesquisadores de todo o mundo, e não somente aos pesquisadores brasileiros, é necessário investigar uma maneira de separar meritocraticamente os autores internacionalmente, ou, caso se verifique uma similaridade de comportamento, verificar a compatibilidade com o modelo proposto.

6. RECONHECIMENTOS

Este trabalho recebeu apoio do InWeb (MCT/CNPq subsídio 57.3871/2008-6) e de subsídios individuais aos autores providos por CNPq, CAPES e FAPEMIG.

REFERENCES

Alpaydin, E. *Introduction to machine learning.* The MIT Press, 2010.

Freire, V. and Figueiredo, D. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, 2011.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.* 20 (4): 422–446, Oct., 2002.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 133–142, 2002.

Joachims, T. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 217–226, 2006.

Newman, M. Who is the best connected scientist? a study of scientific coauthorship networks. *Complex networks*, 2004a.

Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. (PNAS)* 101 (Suppl 1): 5200–5205, 2004b.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6 (2): 1453, 2006.