Uma Abordagem Multi-Visão para a Estimativa de Qualidade de Artigos de Wikis

Daniel Hasan Dalip[1], Thiago Cardoso[1], Marcos André Gonçalves[1], Marco Cristo[2], Pável Calado[3]

[1] Universidade Federal de Minas Gerais
[2] Universidade Federal do Amazonas
[3] Instituto Superior Técnico/INESC-ID
{hasan,thiagon,mgoncalv}@dcc.ufmg.br, marco.cristo@dcc.ufam.edu.br,
pavel.calado@tagus.ist.utl.pt

Abstract. A Wikipédia é um exemplo de repositório de livre acesso e edição criado através do esforço colaborativo de sua comunidade de usuários. Porém, esta enorme quantidade de informação causa uma grande preocupação quanto à qualidade de seu conteúdo, dada a sua disponibilização absolutamente democrática. Para lidar com este problema, alguns trabalhos procuram estimar a qualidade dos artigos na Wikipedia automaticamente. Em vários destes, um grande número de indícios de qualidade é coletado e, em seguida, combinado utilizando técnicas de aprendizado de máquina, com o intuito de se obter um valor único referente à qualidade desses artigos. Cada grupo de indícios pode ser visto como uma diferente visão do conceito de qualidade. Neste trabalho, propomos uma nova abordagem para combinar estas diferentes visões inspirada no método de meta-aprendizado conhecido como empilhamento. Em particular, tomando o conjunto de indícios agrupados em três visões (textual, histórico de revisões e grafo de referências), nós demonstramos que é possível usar esta abordagem em enciclopédias colaborativas como a Wikipedia, obtendo ganhos de até 18% frente ao método de estimativa de qualidade, considerado estado-da-arte.

Categories and Subject Descriptors: H.3.7 [**Information Storage and Retrieval**]: Digital Libraries, User Issues

General Terms: Human Factors, Measurement, Experimentation

Keywords: Avaliação da Qualidade, Wikipedia, Aprendizado de Máquina, SVM, Multi-Visões

1. INTRODUÇÃO

Através da *Web 2.0* e sua natureza colaborativa, um novo tipo de repositório do conhecimento humano está sendo criado. Tais repositórios são caracterizados por serem de livre acesso não apenas para a leitura, como também para a escrita. Este novo meio de propagação do conhecimento é provido, entre outros, por *blogs*, fóruns e bibliotecas digitais colaborativas que fornecem coleções de documentos criados e mantidos pela própria comunidade *web* [Dondio et al. 2006].

Contudo, tal liberdade traz consigo uma importante questão: dada a retórica de acesso democrático a tudo por todos e em qualquer momento, como o usuário pode determinar a qualidade da informação do que ele acessa? Atualmente, o conteúdo gerado de forma centralizada em meios físicos, tais como livros e artigos de revista, ainda é visto mais naturalmente como de qualidade e confiável [Dondio et al. 2006].

Comunidades colaborativas já possuem técnicas manuais para tratar o problema de qualidade dos artigos, levando em conta o julgamento humano. Por exemplo, a comunidade da Wikipédia avalia seus artigos através de seus usuários que os classificam, manualmente, verificando alguns aspectos qualitativos tais como ponto de vista neutro, estrutura do texto, referências bibliográficas, entre outros [Wikipedia 2008]. Entretanto, se considerarmos o tamanho da coleção e a velocidade com que ela se expande, pode se tornar impraticável a avaliação deste conteúdo manualmente [Voß 2005]. Além disso, a avaliação por seres humanos pode ser tendenciosa e influenciada de acordo com a sua cultura, conhecimentos e até mesmo intenções maliciosas [Hu et al. 2007].

Uma possível solução para este problema seria estimar automaticamente a qualidade desses documentos. Com este objetivo, algumas abordagens tem sido propostas na literatura. Dentre estas,

aquela que apresentou os melhores resultados foi sugerida pelos autores em [Dalip et al. 2009]. Nesta abordagem, diversos indícios de qualidade (por exemplo, o tamanho do artigo, o número de revisões recebidas e o número de citações feitas) são combinadas através de uma técnica de aprendizado de máquina. Ao analisar os indícios utilizados nesta e em outras abordagens, notamos que eles podem ser agrupados em textuais, relacionadas com o grafo de referências entre os artigos e relacionadas com o histórico de revisões. Cada um destes grupos pode ser interpretado como uma diferente visão do conceito de qualidade.

A utilização de diversas visões foi motivada pelo fato de que autores [Muslea et al. 2002; Kakade and Foster 2007] têm demonstrado que a combinação de diversas visões pode melhorar o desempenho de métodos de aprendizado de máquina. Como uma visão representa uma diferente percepção de um conceito, a combinação de modelos criados especificamente para cada visão representa melhor a combinação da opinião de diferentes especialistas. Assim, neste trabalho, propomos estimar qualidade a partir de indícios através de (1) organização de tais indícios em diferentes visões e, em seguida, (2) a combinação destas visões usando uma técnica de combinação de modelos de aprendizado. Em particular, para a combinação de modelos, usamos uma técnica inspirada na estratégia de meta-aprendizado conhecida como empilhamento [Wolpert 1992; Blum and Mitchell 1998].

Usando nossa abordagem para combinar as três visões de qualidade descritas em três enciclopédias colaborativas (Wikipédia[1], Starwars[2] e Muppets[3]), obtivemos um ganho de até 18% sobre o trabalho proposto em [Dalip et al. 2009]. Resumindo, nossas contribuições são a proposta de uma nova abordagem de para a combinação de modelos baseados em visões para estimar a qualidade de artigos e a aplicação bem sucedida desta abordagem em 3 enciclopédias colaborativas.

2. TRABALHOS RELACIONADOS

Diversos trabalhos ressaltam o problema da qualidade de bibliotecas colaborativas na *Web*. Uma abrangente revisão destes trabalhos está disponível em [Dalip et al. 2009]. Aqui, nós citamos apenas alguns trabalhos mais significativos.

O primeiro trabalho a sugerir a combinação de vários indícios para estimar qualidade e credibilidade no domínio da Wikipédia foi apresentado em [Dondio and Weber 2006]. Em particular, eles combinaram indícios relacionados com diferentes aspectos de qualidade como estabilidade do artigo, qualidade de edição e importância do artigo. Estes indícios foram extraídos do histórico de revisão, conteúdo do texto e de sua estrutura de ligações. Um ranking final de qualidade foi obtido através de uma combinação linear dos indícios, proposta pelos autores.

Os autores em [Rassbach et al. 2007] propuseram que a combinação dos indícios poderia ser realizada usando uma técnica de aprendizado de máquina. Assim, eles propuseram o uso de um modelo de entropia máxima [Borthwick et al. 1998] para estimar a qualidade dos artigos de acordo com as classes inseridas previamente por avaliadores humanos. Estes autores propuseram vários indícios novos de conteúdo textual para este problema.

Em [Dalip et al. 2009], também foram utilizadas técnicas de aprendizado de máquina para estimar a qualidade dos artigos. Entretanto, utilizou-se regressão baseada em máquina de vetores de suporte (SVR, do original em inglês *Support Vector Regression*) para a tarefa [Vapnik 1995]. A principal contribuição deste trabalho é um estudo detalhado de vários indícios e o seu impacto na previsão da qualidade de um documento, gerado por uma comunidade Web de forma colaborativa. Além disso, o desempenho do método proposto alcançou um resultado superior aos melhores trabalhos publicados anteriormente, tanto em termos da estratégia de aprendizado empregada quanto dos indícios usados.

Todos os métodos citados anteriormente desenvolvem um *único* modelo para combinar os indícios propostos. Contudo, ao observarmos estes indícios, notamos que eles representam três visões distintas

[1] http://en.wikipedia.org/
[2] http://starwars.wikia.com
[3] http://muppet.wikia.com

de qualidade, ou seja, sob um ponto de vista do (a) texto escrito, das (b) revisões realizadas e das (c) conexões entre as páginas. Como mencionado em [Muslea et al. 2002; Blum and Mitchell 1998] múltiplas visões podem ser utilizadas para se ter várias opiniões independentes sobre um processo de classificação. Isto é possível quando há várias formas de classificar um elemento. Por exemplo, um vídeo pode ser classificado com base em seu conteúdo de áudio e vídeo.

Assim, ao contrário de todos estes métodos, pretendemos aprender um modelo para cada visão e, depois, combinar todos os modelos. Para tanto, iremos usar uma técnica de meta-aprendizado baseada em empilhamento [Wolpert 1992]. No empilhamento, um meta-classificador aprende a relação existente entre a saída de diferentes algoritmos de aprendizado e uma classe-alvo. Em nosso caso, em lugar de usar modelos gerados por diferentes algoritmos, iremos usar modelos gerados para diferentes visões. Neste sentido, o método de meta-aprendizado que propomos é ligeiramente diferente do empilhamento.

3. MODELAGEM DO PROBLEMA

Suponha que três especialistas avaliam a qualidade de um artigo, cada qual com uma diferente visão de qualidade (por exemplo, o conteúdo textual, o histórico de revisões e o grafo de ligações). A estimativa final de qualidade deve ser uma combinação destas múltiplas opiniões. Em particular, se cada opinião for dada como um grau de certeza (i.e., um valor referente a qualidade para cada visão), para um certo artigo, é possível aprender sua qualidade global a partir das certezas relacionadas com cada visão. Da mesma forma, a certeza relacionada com cada visão pode ser aprendida a partir dos vários indícios que constituem a visão.

Logo, o problema de estimar qualidade pode se dar em duas fases de aprendizado. Na primeira fase, ou aprendizado de nível 0, cada artigo é representado por um conjunto de indícios relacionados com uma visão particular. Assim, cada artigo possui três representações (i.e., conjunto de atributos), uma para cada visão. Um modelo de qualidade é aprendido para cada visão. Como resultado, para cada artigo, temos uma estimativa de qualidade relacionada com cada visão. Em uma segunda fase, ou aprendizado de nível 1, cada artigo é representado pelo conjunto de estimativas de qualidade de cada visão. Um modelo de qualidade global é aprendido e, como resultado, para cada artigo, temos uma estimativa final de qualidade. Nas próximas seções, estas representações são discutidas em detalhes.

3.1 Aprendizado de nível 0

Na Wikipédia, a qualidade de um artigo é atribuída como um valor em uma escala discreta. Assim, existem os resumos, os esboços iniciais, os artigos de classe B, os bons artigos, os de classe A e os artigos de destaque. Note, entretanto, que qualidade, de forma geral, pode ser vista como um valor em uma escala contínua, variando do pior para o melhor. De fato, essa é uma interpretação mais natural para o problema se considerarmos que há artigos melhores e piores mesmo dentro de uma mesma classe discreta. Por exemplo, no caso da Wikipédia, temos artigos de classe A que (a) acabaram de ser promovidos e esperam pela avaliação de especialistas, (b) que já foram avaliados por especialistas e esperam correções e (c) que já foram corrigidos e esperam promoção para a classe de destaque. No caso de outras *Wikis*, geralmente é utilizado uma escala contínua para a avaliação manual de seus artigos onde os usuários pontuam cada artigo de 1 a 5 e a nota resultante de um artigo é a média de sua pontuação.

Por estes motivos, neste trabalho, iremos considerar qualidade como uma escala contínua e, consequentemente, o problema de aprender esta escala em nível 0 será modelado como uma tarefa de regressão. Em particular, aplicaremos um método estado-da-arte para o aprendizado de regressão, o Support Vector Regression (SVR).

3.1.1 *Estimativa de qualidade com SVR.* Para utilizar o SVR na tarefa de estimar a qualidade de artigos, representamos os artigos como descrito a seguir. Dada uma certa visão v de qualidade, seja $A_v = \{a_{v1}, a_{v2}, ..., a_{vn}\}$ um conjunto de artigos, onde cada artigo a_{vi} é representado por conjunto de m atributos $\{F_{v1}, F_{v2}, ..., F_{vm}\}$. Assim, $a_{vi} = (f_{vi1}, f_{vi2}, ..., f_{vim})$, onde f_{vij} é o valor do atributo F_{vj}

no artigo a_{vi}. Neste nível, um *atributo* representa um indicador de qualidade de um artigo de acordo com a visão v. Por exemplo, f_{vij} poderia representar o tamanho do artigo a_{vi} em uma visão textual de qualidade.

Assuma que possuímos um conjunto de treino $A_v \times \mathbb{R} = \{(a_{v1}, q_{v1}), (a_{v2}, q_{v2}), ..., (a_{vn}, q_{vn})\}$, onde cada par (a_{vi}, q_{vi}) representa o artigo a_{vi} e sua qualidade q_{vi}, de acordo com a visão v. Assim, dada a visão v, se $q_{v1} > q_{v2}$, a qualidade do artigo a_{v1}, segundo o usuário, é maior que a qualidade do artigo a_{v2}.

A solução proposta para este problema consiste em: (1) determinar o conjunto das visões v; (2) determinar o conjunto de atributos $\{F_{v1}, F_{v2}, ..., F_{vm}\}$ utilizados para representar os artigos em A_v e (3) aplicar o método de aprendizado de regressão (SVR) para encontrar a melhor combinação de atributos para estimar a qualidade q_{vi} de um artigo a_{vi} para cada visão v.

O objetivo do SVR é achar uma função $\gamma : A_v \rightarrow \mathbb{R}$ que possua um erro de no mínimo ϵ dos resultados obtidos r_i, em todo o conjunto de treino, e que o erro seja minimizado de uma forma balanceada.

Dada a função $\gamma(x) = \kappa(w, x) + b$, onde κ representa a função de produto interno (ou função de núcleo) em um determinado espaço, w representa o vetor de m atributos, $b \in \mathbb{R}$ é uma constante e x representa a instância de documento ou item que possuirá seu resultado alvo r estimado. O processo de regressão consiste em aprender w e b a partir do conjunto de treino. Para tornar γ com o menor erro possível e de forma balanceada, é necessário minimizar w ($||w||$).

Formalmente, pode-se definir isto como um problema de otimização quadrática minimizando:

$$\frac{1}{2} \|\overrightarrow{w}\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \tag{1}$$

Sujeito a:

$$\begin{aligned} q_{vi} - \kappa(\overrightarrow{w}, \overrightarrow{x}_i) - b &\leq \epsilon + \xi_i \\ \kappa(\overrightarrow{w}, \overrightarrow{x}_i) + b - q_{vi} &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

onde ξ_i e ξ_i^* são variáveis criadas para aumentar a tolerância ao erro na otimização. A constante $C > 0$ é definido para ponderar a importância entre os erros acima de ϵ tolerados e a minimização de γ.

Este é um problema de otimização quadrática que, neste trabalho, foi solucionado utilizando o programa SVMLIB [Chang and Lin 2001]. Além disso, foi utilizado uma função de base radial [Witten and Frank 1999] (RBF, do original em inglês *Radial Basis Function*) como κ. Na próximas seções, serão apresentados as formas de combinações utilizadas e os atributos utilizados para representar os artigos.

3.1.2 *Representações do Artigo.* Para estimar a qualidade de artigos utilizando regressão é essencial determinar quais atributos serão utilizados. Tais atributos são baseados nos critérios utilizados pela Wikipédia [Wikipedia 2008], para qualificar um artigo. Segundo a Wikipédia, um bom artigo precisa ser compreensível, bem estruturado e organizado, ser completo e não prolixo, possuir referências bibliográficas e uma visão neutra. Estes atributos foram divididos em 3 visões: textuais, histórico de revisões e ligações.

Atributos textuais foram extraídos do conteúdo textual dos artigos. Eles podem ser divididos em quatro sub-grupos: tamanho, estilo, estrutura e legibilidade. São exemplos destes atributos o número de caracteres do artigo (tamanho), o percentual de frases longas (estilo), a distribuição das seções (estrutura) e a métrica de legibilidade *Flesch Reading Ease* [Flesch 1948], que estima o nível de instrução que uma pessoa deve possuir para conseguir compreender um texto.

Os atributos extraídos do histórico de revisão são úteis para estimar a maturidade e estabilidade do texto [Dondio and Weber 2006]. Espera-se que um artigo atinja um nível de maturidade suficiente

para não necessitar de grandes alterações tornando-se assim, estável. São exemplos destes atributos a idade, o numero de revisões por dia, e a métrica ProbReview [Hu et al. 2007], que estima a qualidade do artigo baseada na qualidade de seus autores.

Atributos de ligações são extraídos através do grafo de artigos existente na coleção onde os vértices são artigos da Wikipédia e as arestas são os apontadores entre um artigo e outro. São exemplos de atributos de ligações o coeficiente de clusterização de um artigo [Dorogovtsev and Mendes 2003], o seu PageRank [Brin and Page 1998], o número de referencias feitas por ele e o número de vezes em que ele é referenciado.

Neste trabalho, utilizamos todos os 68 atributos descritos em [Dalip et al. 2009]. Dado o espaço limitado que dispomos, sugerimos que o leitor interessado em uma explicação detalhada dos atributos, leia diretamente o texto em [Dalip et al. 2009].

3.2 Aprendizado de nível 1

Uma vez que a qualidade dos artigos foi prevista para cada visão, uma nova representação pode ser usada para descrevê-los. Assim, cada artigo a_i é representado por um conjunto de três atributos $\{F_1, F_2, F_3\}$. Neste nível, cada *atributo* representa a estimativa de qualidade do artigo de acordo com uma diferente visão. Assim, F_1 representa a qualidade do artigo de acordo com uma visão textual, F_2 a qualidade de acordo com o histórico e F_3 a qualidade de acordo com as ligações. Dado um conjunto de treino $\{(a_1, q_1), (a_2, q_2), ..., (a_n, q_n)\}$, onde cada par (a_i, q_i) representa o artigo a_i e sua qualidade final q_i, a qualidade de um novo artigo pode ser aprendida por aplicarmos SVR como descrito na seção anterior. Note que ao fazermos isto estamos, de fato, aprendendo a combinar as estimativas obtidas para cada diferente visão de qualidade.

4. EXPERIMENTOS

Utilizando todos os atributos da Seção 3.1.2 foram realizados uma série de experimentos. Neste capítulo será apresentado a metodologia adotada nos experimentos e seus resultados.

4.1 Coleção Utilizada

Para nosso experimentos, decidimos utilizar uma amostra da Wikipédia em inglês e amostra de duas enciclopédias colaborativas da *Wikia*. Utilizamos artigos da Wikipédia em inglês devido ao tamanho da coleção, que atualmente possui 3.1 milhões de artigos sendo que aproximadamente 1.680.000 destes artigos já foram avaliados pelo usuário quanto à qualidade [Wikipedia 2011]. Além disso, a Wikipédia possui o seu repositório disponível para download possibilitando assim a extração dos artigos e seus atributos[4]. Esta coleção será chamada a partir de agora de WIKIPEDIA. A avaliação do artigo da Wikipédia pode ser feita por qualquer usuário, obedecendo a seguinte escala de qualidade[5] [Wikipedia 2011]:

—*Featured Article* (*FA*): Em português "Artigo em destaque". Estes são, de acordo com os avaliadores, os melhores artigos da Wikipédia.

—*A-Class* (*AC*): Estes artigos são considerados completos, porém ainda com pequenas pendências a serem solucionadas como a correção da formatação utilizada.

—*Good Article* (*GA*): Em português "bom artigo". São artigos sem problemas de lacunas ou conteúdo excessivo. Eles são boas fontes de informação, porém outras enciclopédias podem fornecer um material melhor.

—*B-Class* (*BC*): artigos úteis para a maioria dos usuários. Especialistas, entretanto, podem necessitar de informações mais precisas.

[4] `http://en.wikipedia.org/wiki/Wikipedia_database`

[5] Atualmente, há uma classe nova entre ST e BC, a *C-Class*, porém, esta classe não existia quando foram coletados os artigos.

—*Start-Class* (*ST*): Artigo ainda incompleto, porém contendo alguma referência para que se possa obter uma informação mais completa.

—*Stub-Class* (*SB*): Estes são artigos rascunho. Geralmente consistem de poucos parágrafos de texto com nenhuma ou pouca estrutura.

Do serviço Wikia, escolhemos duas enciclopédias colaborativas a *Wookieepedia*[6] (Enciclopédia sobre Starwars) e a *Muppet*[7] (Enciclopédia sobre o seriado infantil *The Muppet Show*), respectivamente. Utilizamos estas duas enciclopédias, pois, dentre as enciclopédias colaborativas da wikia, estas são as que possuem mais artigos em cada classe de qualidade[8]. Além disso, estas duas coleções possuem também seu repositório disponível para download[9] e o formato do repositório é o mesmo da Wikipédia, facilitando sua extração.

A coleção Wookieepedia fornece dois tipos de taxonomias para a qualidade. A primeira é um subconjunto da taxonomia da Wikipédia formada pelas classes FA, GA e SB. A segunda é baseada na taxonomia que é geralmente usada nas coleções da Wikia: uma taxonomia utilizando estrelas, onde o usuário define se o artigo é de 1 até 5 estrelas através de notas, sendo que 1 é a mais baixa qualidade para 5 que é a mais alta. Desta forma, a nota do artigo é a média da nota dos usuários. Como essas duas taxonomias não são compatíveis, criamos duas amostras a STWR_3CLASS, baseada na taxonomia da Wikipédia, e STWR_5CLASS, baseada na taxonomia utilizando estrelas. Na coleção Muppet, há apenas a taxonomia baseada em estrelas que chamaremos a amostra de MUPPET.

O tamanho de cada amostra é apresentado na Tabela 4.1. Cada amostra foi criada de forma balanceada, onde foi resgatado 100% dos artigos da classe com o menor número de artigos e, logo após, extraindo o mesmo número de artigos aleatoriamente para todas as outras classes.

Os atributos de ligações foram obtidos através dos apontadores entre as páginas internas de cada coleção. Estes apontadores foram extraídos através de um arquivo de importação disponível para download[10]. A quantidade de vértices (artigos e páginas que redirecionam para artigos) e arestas (apontadores de uma pagina para outra) de cada coleção é apresentado na Tabela 4.1. Foi utilizado o programa denominado *Web Graph* [Boldi and Vigna 2004] para a compactação do grafo e cálculo dos atributos de redes complexas.

Amostra	# Artigos	# Revisões	# Arestas	# Vértices	Data da coleção
WIKIPEDIA	3.294	1.992.463	86.077.675	3.185.457	janeiro/2008
MUPPET	1.550	38.291	282.568	29.868	setembro/2009
STWR_3CLASS	1.446	127.551	1.017.241	106.434	outubro/2009
STWR_5CLASS	9.180	369.785	1.017.241	106.434	outubro/2009

Table I. Tamanho da amostra

4.2 Metodologia de Avaliação

Os experimentos realizados tiveram o objetivo de realizar uma analise comparativa entre métodos de combinação de visões.

Como foi proposto neste trabalho um método baseado em regressão, utilizou-se uma escala de qualidade, onde cada classe foi transformada em número variando de 0 (*Stub article* ou 1 estrela) a 5 (*Featured Article* ou 5 estrelas). Sendo assim, Para verificarmos o desempenho do método de regressão

[6]http://starwars.wikia.com/
[7]http://muppet.wikia.com/
[8]Para a extração da avaliação dos artigos, foi utilizado a API de cada Wiki disponível em: http://starwars.wikia.com/api.php e http://muppet.wikia.com/api.php
[9]http://starwars.wikia.com/wiki/Special:Statistics e http://muppet.wikia.com/wiki/Special:Statistics
[10]Na Wikipédia: `http://en.wikipedia.org/wiki/Wikipedia_database`, nas demais coleções o grafo foi extraído através da estrutura de links do conteúdo da página

proposto, utilizamos o erro quadrático médio (MSE, do original em inglês *Mean Squared Error*) que é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e^2 \tag{2}$$

onde e é o valor do erro e n é o tamanho da amostra.

Ao avaliar a qualidade de um artigo utilizando regressão, avaliadores humanos pontuariam, com valores estipulados em uma escala, os artigos quanto à sua qualidade e o erro e seria a distância do valor de qualidade estimado pela regressão e o valor atribuído por estes avaliadores.

Nos métodos de aprendizado de máquina apresentados neste trabalho, os experimentos foram realizados utilizando validação cruzada de dez partições (10-fold cross validation) [Mitchell 1997], tanto no aprendizado de nível 0 quanto no aprendizado de nível 1. Desta forma, a coleção foi dividida aleatoriamente em dez partes e cada experimento foi repetido 10 vezes. Em cada repetição, uma partição diferente passou a ser o teste, enquanto o restante foi usado para o treino. As partições utilizadas para treino e teste foram as mesmas em todos os experimentos. O resultado final de cada experimento representa a média destas 10 rodadas. Note, entretanto, que diferentes particionamentos são usados em cada nível, da mesma forma que em [Wolpert 1992].

Durante as comparações realizadas neste trabalho, com o objetivo de determinar se a diferença de desempenho é estatisticamente significativa, utilizou-se o teste de postos com sinais de Wilcoxon (do inglês, Wilcoxon *signed-rank test*) [Wilcoxon 1945]. Em todos os casos, apenas tiramos conclusões dos resultados que foram estatisticamente significativos com, no mínimo, 90% de confiança.

Para garantir que os resultados não foram afetados por uma escolha inadequada de parâmetros, vários experimentos foram realizados e, em todos os casos, relatamos apenas os melhores resultados.

4.3 Resultados

Tabela II demonstra o resultado dos experimentos para cada coleção. A partir de agora chamaremos de SVR o método *baseline* proposto por [Dalip et al. 2009], MA_VIEW seria o método de meta-aprendizado utilizando apenas as features de visões no aprendizado de nível 1 e MA_VIEW_ARTICLE foi utilizado no aprendizado de nível 1, além das features que representam o resultado de cada visão, as 68 features que representam o artigo (união das 3 visões). Além disso, valores de MSE marcados com '*' na Tabela II indicam diferenças estatisticamente significativas em comparação ao SVR.

Amostra	Método	MSE	% melhoria
WIKIPEDIA	SVR	0,856	-
	MA_VIEW	0,84*	1,02 %
	MA_VIEW_ARTICLE	0,809*	5,8 %
MUPPET	SVR	1,685	-
	MA_VIEW	1,676	0,5 %
	MA_VIEW_ARTICLE	1,682	0,2 %
STWR_5CLASS	SVR	1,681	-
	MA_VIEW	1,665*	0,9 %
	MA_VIEW_ARTICLE	1,661*	1,2 %
STWR_3CLASS	SVR	0,075	-
	MA_VIEW	0,061*	18,6 %
	MA_VIEW_ARTICLE	0,067*	10,6 %

Table II. Erro quadrático médio por método em cada amostra

Como foi possível observar, o meta-aprendizado conseguiu melhorar o resultado em todas as amostras exceto na MUPPET, além disso, quando foi utilizado informação a respeito do artigo houve uma melhoria maior na predição na amostra WIKIPEDIA. Dessa forma foi possível observar que o meta-aprendizado pode ser útil para estimativa da qualidade de artigos em várias coleções wikis. E a importância de utilizar a representação de artigos para a combinação de visão nesta tarefa.

5. CONCLUSÃO

Neste trabalho, foi utilizado a proposta de [Dalip et al. 2009], em um contexto mais amplo, em 3 bibliotecas colaborativas com características diferentes. Para cada biblioteca colaborativa, foram extraídos atributos para representar cada artigo e, logo após, estes atributos foram divididos em 3 visões com o objetivo de utilizar métodos de combinação de visões e melhorar a precisão em comparação ao trabalho de [Dalip et al. 2009]. Conseguimos diminuir o erro da previsão em todas as encilopédias colaborativas testadas, exceto na Muppets.

Como trabalho futuro, pretendemos explorar e comparar outros métodos de combinação de visão, analisar o desempenho ao reduzir o treino utilizando a representação multi-visão do problema. Finalmente, pretendemos criar ferramentas para auxiliar no processo de classificação de qualidade de enciclopédias, bem como estudar a influência deste fator em processos de busca.

Agradecimentos. Este trabalho foi parcialmente financiado pelo InWeb (MCT / CNPq 57.3871/2008-6) e através de grants e bolsas de estudo do CNPq, CAPES e FAPEMIG individuais dos autores.

REFERENCES

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. COLT' 98. ACM, New York, NY, USA, pp. 92–100, 1998.

Boldi, P. and Vigna, S. The webgraph framework I: Compression techniques. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*. ACM, New York, NY, USA, pp. 595–601, 2004.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the Sixth Workshop on Very Large Corpora*, 1998.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30 (1-7): 107–117, April, 1998.

Chang, C. C. and Lin, C. J. LIBSVM: a library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *JCDL '09: Proceedings of the 2009 Joint International Conference on Digital libraries*. Austin, TX, USA, pp. 295–304, 2009.

Dondio, P., Barrett, S., Weber, S., and Seigneur, J. Extracting trust from domain analysis: A case study on the wikipedia project. In *Autonomic and Trusted Computing*. Springer Berlin / Heidelberg, pp. 362–373, 2006.

Dondio, Pierpaolo, S. B. and Weber, S. Calculating the trustworthiness of a wikipedia article using dante methodology. In *IADIS International Conference on e-Society*. Dublin, Ireland, 2006.

Dorogovtsev, S. N. and Mendes, J. F. F. *Evolution of Networks: From Biological Nets to the Internet and WWW (Physics)*. Oxford University Press, 2003.

Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. Lisbon, Portuga, pp. 243–252, 2007.

Kakade, S. M. and Foster, D. P. Multi-view regression via canonical correlation analysis. In *In Proc. of Conference on Learning Theory*, 2007.

Mitchell, T. M. *Machine Learning*. McGraw-Hill Higher Education, 1997.

Muslea, I., Minton, S., and Knoblock, C. A. Active semi-supervised learning = robust multi-view learning. In *Proceedings of ICML-02, 19th International Conference on Machine Learning*. pp. 435–442, 2002.

Rassbach, L., Pincock, T., and Mingus, B. Exploring the feasibility of automatically rating online article quality. `http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf`, 2007.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

Voss, J. Measuring wikipedia. In *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*. Number PREPRINT 2005-04-12. Karolinska University Press, 2005.

Wikipedia. Version 1.0 editorial team/release version criteria. `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Release_Version_Criteria`, 2008.

Wikipedia. Version 1.0 editorial team/assessment. `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`, 2011.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics*, 1945.

Witten, I. H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 1999.

Wolpert, D. H. Stacked generalization. *Neural Networks* vol. 5, pp. 241–259, 1992.