UM MÉTODO AUTOMÁTICO PARA ESTIMATIVA DA QUALIDADE DE ENCICLOPÉDIAS COLABORATIVAS ON-LINE: UM ESTUDO DE CASO SOBRE A WIKIPÉDIA

DANIEL HASAN DALIP

UM MÉTODO AUTOMÁTICO PARA ESTIMATIVA DA QUALIDADE DE ENCICLOPÉDIAS COLABORATIVAS ON-LINE: UM ESTUDO DE CASO SOBRE A WIKIPÉDIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Marcos André Gonçalves
Co−Orientador: Marco Cristo

Belo Horizonte

Março de 2009

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Um Método Automático para Estimativa da Qualidade de Enciclopédias Colaborativas On-Line: Um Estudo de Caso Sobre a Wikipédia

DANIEL HASAN DALIP

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARCOS ANDRÉ GONÇALVES - Orientador
Departamento de Ciência da Computação - UFMG

PROF. MARCO ANTÔNIO PINHEIRO DE CRISTO - Co-orientador
Fundação Centro de Análise Pesquisa e Inovação Tecnológica - FUCAPI

PROFA. RENATA DE MATOS GALANTE
Instituto de Informática - UFRGS

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 03 de abril de 2009.

Agradecimentos

Agradeço a todos que direta ou indiretamente ajudaram a realizar o meu trabalho, em especial:

- A minha família pelo total apoio
- Aos meus orientadores Marcos André Gonçalves e Marco Cristo pela orientação, apoio e dedicação
- À Gisele L. Pappa e Renata Galante pelas sugestões
- À todos os meus amigos que me ajudaram de alguma forma nesta jornada, em especial Maria Raquel, Mariana, Luisa Rocha e todos os integrantes e amigos da banda Mutum
- Aos meus novos amigos do mestrado pelo apoio e ajuda em todas as ocasiões: Emiliana, Sandra e Lucas
- À todos que torceram por mim e me apoiaram em especial: Maria Raquel, Mariana, Juliane, Mari (cotinha), Livia, Livia Mara, Sandra, professores do curso de computação do Uni-BH e meus colegas de trabalho do Uni-BH
- Ao corpo docente do curso de Ciência da Computação do Centro Universitário de Belo Horizonte (Uni-BH)
- Ao corpo docente da UFMG
- À Magali Araujo, minha orientadora de graduação no Uni-BH, pelo incentivo, orientação e amizade
- À toda equipe da Coordenadoria de Sistemas do Centro Universitário de Belo Horizonte pelo apoio e amizade
- Aos colegas do Laboratório de Banco de Dados

"If you want to have cities, you've got to build roads"
(Cake)

Resumo

O antigo sonho de livre acesso a um repositório contendo todo o conhecimento e cultura humana está se tornando realidade através da Internet e da participação colaborativa dos seus usuários. A Wikipédia é um grande exemplo de repositório de livre acesso e edição criado através do esforço colaborativo de sua comunidade de usuários. Entretanto, esta enorme quantidade de informação disponibilizada de forma democrática causa uma grande preocupação quanto à qualidade de seu conteúdo. Nesta dissertação foram coletados um grande número de indícios de qualidade, alguns já propostos e outros novos e utilizados pela primeira vez para este fim, com o objetivo de estudar a capacidade dos mesmos em estimar a qualidade dos artigos da Wikipédia. Além disso, foi apresentada uma nova abordagem para combinar estes indícios, utilizando técnicas de aprendizado de máquina, para extrair um valor unificado referente à qualidade desses artigos. Com este trabalho foi possível analisar o impacto de cada indício e foi identificado, por exemplo, que os mais promissores em estimar a qualidade de um artigo são aqueles mais simples de extrair em qualquer biblioteca digital de acesso livre: os que levam em conta o conteúdo do texto atual como, por exemplo, o tamanho do artigo e a sua estrutura. Também foi identificado quais indícios não foram tão importantes na estimativa da qualidade. Estes indícios são, coincidentemente, os mais complexos de extrair, como os baseados em análise de ligações. Por fim, o método proposto demonstrou ganhos significativos na estimativa da qualidade de artigos ao compará-lo com as soluções estado-da-arte na literatura.

Abstract

The old dream of a universal repository containing all the human knowledge and culture is becoming possible through the Internet and the Web. Moreover, this is happening with the direct collaborative, participation of people. Wikipedia is a great example. It is an enormous repository of information with free access and edition, created by the community in a collaborative manner. However, this large amount of information, made available democratically and virtually without any control, raises questions about its relative quality. In this work we explore a significant number of quality indicators, some of them proposed by us and used here for the first time, and study their capability to assess the quality of Wikipedia articles. Furthermore, we explore machine learning techniques to combine these quality indicators into one single assessment judgment. Through experiments, we show that the most important quality indicators are the easiest ones to extract on a open digital library, namely, textual features related to length, structure and style. We were also able to determine which indicators did not contribute significantly to the quality assessment. These were, coincidentally, the most complex features, such as those based on link analysis. Finally, we compare our combination method with state-of-the-art solutions and show significant improvements in terms of effective quality prediction.

Keywords: Quality Assessment,Wikipedia, Machine Learning, SVM.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Através da Internet, 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, mas também para a escrita. Este novo meio de propagação do conhecimento é provido, entre outros, por blogs e bibliotecas digitais colaborativas que fornecem coleções de documentos criados e mantidos pela própria comunidade web [Krowne, 2003; Dondio et al., 2006].

Uma amostra de como comunidades podem produzir conteúdo colaborativo rapidamente e em larga escala é dada pela Wikipedia[1]. Esta enciclopédia online colaborativa levou apenas dois anos para acumular tantos artigos quanto a Enciclopédia Britânica e conta, atualmente, com mais de sete milhões de artigos escritos em centenas idiomas [Wikipedia, 2008a]. Porém, a Wikipédia é apenas um exemplo. Atualmente, há mais de cinqüenta serviços de hospedagem de coleções que permitem a livre edição do seu conteúdo por seus usuários [Fogg et al., 2003; Dondio et al., 2006]. Cada serviço destes conta com inúmeras coleções abordando temas variados como: informações geográficas, esportes, tecnologia, ciência, programação de TV, ficção, eventos, livros, conhecimentos gerais, entre outros. O mais popular destes serviços de hospedagem, a *Wikia*[2], cresceu de cem para mais de três mil coleções em apenas dois anos e atualmente possui algumas coleções contendo mais de cinqüenta mil documentos [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

[1] http://www.wikipedia.org/
[2] http://www.wikia.org

é visto mais naturalmente como de qualidade e confiável [Dondio et al., 2006]. Entretanto, dada a disseminação deste conteúdo colaborativo de livre edição, a confiança neste tipo de material terá de ser maior.

Uma possível solução para este problema é desenvolver um sistema para estimar a qualidade desses artigos automaticamente. Eles poderiam ser usados para indicar que artigos deveriam ser revisados, detectar processos de revisão inadequados tais como vandalismos ou recomendar diretamente o material aos usuários baseado nos seus indícios de qualidade e confiabilidade.

Sendo assim, neste trabalho é apresentado uma nova forma de combinar indícios, baseada em regressão e aprendizado de máquina, com o objetivo de extrair um valor único referente à qualidade do artigo. Vários desses indícios já foram propostos em outros trabalhos, enquanto outros são novos. Realizamos um estudo detalhado do impacto destes indícios ao estimar a qualidade dos artigos. Além disso, ao compararmos nosso trabalho com outros previamente propostos, obtivemos resultados significativamente superiores na tarefa de estimar a qualidade dos artigos da Wikipédia.

1.1 Definição do Problema

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, 2008e]. Entretanto, se considerarmos o tamanho da coleção e a velocidade com que ela se expande [Voß, 2005], pode se tornar impraticável a avaliação deste conteúdo manualmente. 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].

Desta forma, o problema deste trabalho é combinar vários indícios de um artigo de forma automática retornando, dentro de uma escala pré-determinada, um valor único que represente a sua qualidade.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver uma nova abordagem para estimar a qualidade dos artigos de comunidades colaborativas, levando em conta os indícios mais eficazes já apresentados na literatura

e ainda a proposta de novos indícios.

1.2.2 Objetivos Específicos

1. Identificar indícios já propostos na literatura para este fim;
2. Identificar novos indícios que podem influenciar na percepção dos usuários sobre a qualidade dos artigos de uma comunidade colaborativa;
3. Propor um método, baseado em aprendizado de máquina, para combinar estes indícios e compará-lo com o estado-da-arte;
4. Determinar a melhor combinação e o impacto dos indícios apresentados, levando em consideração o custo computacional para pré-processar cada indício e sua eficácia.

1.3 Contribuições

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. Para tanto, propomos a aplicação de um método, baseada em máquinas de vetores de suporte e regressão, para combinar os indícios. Dado o melhor do nosso conhecimento, esta abordagem é inédita e explora tanto indícios já propostos na literatura quanto outros exclusivos deste trabalho. Além disso, o desempenho do método proposto alcançou um resultado superior aos melhores trabalhos publicados na literatura, tanto em termos da estratégia de aprendizado empregada quanto dos indícios usados. As contribuições dessa dissertação foram formalmente apresentadas em Dalip et al. [2009].

1.4 Trabalhos Relacionados

Diversos trabalhos ressaltam a qualidade dos artigos presentes na Wikipédia. Brown [2009] justifica a boa qualidade dos artigos da Wikipédia comparando-a com a teoria do evolucionismo, onde as revisões seriam mutações e as ruins durariam poucas "gerações", pois existiria uma "seleção natural" das boas revisões. Este autor cita Giles [2005] onde é feito uma comparação mostrando que a precisão da Wikipédia é semelhante a da Enciclopédia Britânica. Entretanto, o fraco controle editorial da Wikipédia possibilitou

que uma biografia falsa de um conhecido jornalista americano permanecesse sem correção por quatro meses, o que motivou uma grande controvérsia e revisão das políticas editoriais do serviço [P. Dondio & Weber, 2006].

A necessidade de determinar a qualidade de conteúdo na Internet tem motivado vários trabalhos na literatura. Por exemplo, Veltman [2005], sugere que a Internet, no futuro, deverá prover mecanismos para lidar com múltiplas variantes de um conteúdo, graus de certeza e importância de afirmações. Um exemplo de tal mecanismo é proposto por Chu [1997], através do qual seria possível calcular a credibilidade de uma afirmação baseado em suas fontes e autores.

Tais soluções, entretanto, pressupõem que a informação necessária será fornecida por autores e/ou usuários. Considerando a natureza livre da Internet, este pode ser um requisito difícil de garantir e implementar. Além disso, a implementação destes requisitos pode ser até mesmo indesejável devido a questões de privacidade e segurança. Todas estas dificuldades estimulam o desenvolvimento de estratégias que busquem estimar a qualidade e a credibilidade de conteúdo online considerando a Internet como ela é hoje, ou seja, utilizando somente os indícios atualmente disponíveis. Exemplos destes indícios, anteriormente exploradas, são ligações e relacionamentos de autoria, forma, identificação dos autores, objetividade, cobertura, correção da informação, ausência de erros tipográficos, indicação de referências e do equilíbrio entre as idéias discutidas além do histórico de avaliações [Cassel, 1995; Alexander & Tate, 1999; Fogg et al., 2001; Roberts, 2004; Stvilia et al., 2005; Korfiatis et al., 2006; Adler & de Alfaro, 2007; Hu et al., 2007].

Com base em tais estudos, P. Dondio & Weber [2006]; Dondio et al. [2006] sugerem uma metodologia para estimar qualidade e credibilidade no domínio da Wikipédia. Estes autores focaram em atribuir um valor indicando a qualidade dos artigos através da combinação de vários indícios. Eles perceberam que peculiaridades desta coleção, como a atualização constante do seu conteúdo e a independência da versão atual em relação às versões anteriores, dificultam a aplicação da maioria dos métodos propostos anteriormente. P. Dondio & Weber [2006] combinam indícios para a construções de rankings que atendam vários aspectos de qualidade (estabilidade do artigo, qualidade de edição, importância do artigo). Estes indícios são extraídos do histórico de revisão, conteúdo do texto e de sua estrutura de ligações. Logo após, estes autores combinaram todos estes rankings para a criação de um ranking único.

Os autores em Rassbach et al. [2007] propuseram uma técnica de aprendizado de máquina para o problema. Esta consistiu no uso do 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

alguns indícios de conteúdo textual para este problema.

No presente trabalho, da mesma forma que Rassbach et al. [2007], também foi utilizado técnicas de aprendizado de máquina para este problema. Entretanto, utilizou-se regressão baseada em máquina de vetores de suporte (SVR, do original em inglês *Support Vector Regression*) [Vapnik, 1995] para estimar a qualidade dos artigos. Além disso, diferente dos trabalhos anteriormente realizados, avaliou-se o impacto tanto dos indícios que já foram propostos na literatura quanto dos novos apresentados neste trabalho.

1.5 Organização do trabalho

Este trabalho está organizado da seguinte forma. No Capítulo 2, é apresentado o embasamento teórico utilizado para implementação e avaliação do método proposto. A metodologia proposta e os atributos utilizados são apresentados no Capítulo 3. No Capítulo 4, são delineados os experimentos realizados e a metodologia adotada para realizá-los. As conclusões e trabalhos futuros são descritos no Capítulo 5.

Capítulo 2

Fundamentação Teórica

Neste capítulo, será apresentado o embasamento teórico utilizado para estimar a qualidade dos artigos e avaliar os experimentos realizados.

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.

Além disso, uma interpretação de qualidade como uma escala contínua é vantajosa em aplicações que uma distinção entre artigos de uma mesma classe é desejável. Este é o caso de um sistema que sugere quais artigos devem ser avaliados por especialistas humanos. Aqueles mais próximos de um certo nível de qualidade seriam os prioritários. Escalas contínuas também seriam mais úteis em sistemas de combinação de *rankings*, por proporcionar uma curva de valores mais suave para todo o conjunto de artigos e mais facilmente interpretável como uma escala probabilística. Finalmente, para um usuário humano tendo que escolher qual versão considerar entre duas alternativas (por exemplo, duas traduções de um mesmo artigo), seria interessante a visualização de um valor numérico que distinguisse as versões, mesmo que ambas pertencessem à mesma classe de qualidade.

Por estes motivos, neste trabalho, iremos considerar qualidade como uma escala contínua e, consequentemente, o problema de aprender esta escala 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). Para avaliarmos os resultados, usaremos duas métricas. A primeira é aquela tipicamente usada em tarefas de regressão, o erro quadrático médio. A segunda, é uma comumente usada para comparação de *rankings*, o NDCG@k. Finalmente, para avaliar o impacto dos atributos, usaremos uma métrica comumente empregada para este fim, o ganho de informação. Tais tarefas, métodos e métricas são descritos nas seções a seguir.

2.1 Regressão

Técnicas de aprendizado de máquina tentam aprender, através de padrões existentes em um conjunto de dados, uma forma de atingir um determinado resultado alvo com o menor erro possível. Este resultado alvo pode ser nominal ou numérico. Quando o valor é numérico e real pode-se utilizar a regressão [Witten & Frank, 1999].

A utilização de regressão para este fim é feita da seguinte forma. Dado o conjunto de treino $A = \{(a_1, r_1), (a_2, r_2), (a_3, r_3), ..., (a_{n-1}, r_{n-1}), (a_n, r_n)\}$, onde a_i possui um ou mais atributos e um resultado alvo r_i, a regressão aprende uma forma de combinar e ponderar os atributos de a_i, de tal forma que o resultado se aproxime o máximo possível do resultado alvo r_i. Sendo assim, após o aprendizado, esta técnica seria capaz de prever o resultado de outras instâncias do conjunto A cujo o resultado alvo é desconhecido.

Neste trabalho, a técnica de aprendizado de máquina utilizada foi regressão baseada em máquinas de vetores de suporte (SVR, do original em inglês *Support Vector Regression*[Vapnik, 1995]), detalhada na próxima seção.

2.1.1 SVR

Supondo que se possui acesso a um conjunto de treino A, definido na seção anterior, o objetivo do SVR é achar uma função $\gamma : A \rightarrow \mathbb{R}$ que possua um erro de, no mínimo, ϵ dos resultados obtidos r_i, em todo o conjunto de treino, e que γ seja tal que se defina um 'tubo' imaginário envolvendo todos os pontos de A, e que seja o mais fino possível.

Seja a função $\gamma(a) = \kappa(w, a)+b$, onde κ representa a função de produto interno (ou função de núcleo) em um determinado espaço, w representa um vetor de m atributos, $b \in \mathbb{R}$ é uma constante e a 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 D. Para tornar o tubo definido por γ o mais fino possível, é necessário minimizar w ($||w||$).

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

$$\frac{1}{2}\|\overrightarrow{w}\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \tag{2.1}$$

Sujeito a:

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

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

Este é um problema de otimização que, no presente trabalho, foi solucionado utilizando o programa SVMLIB [Chang & Lin, 2001]. Neste trabalho, foi utilizado uma função de base radial [Witten & Frank, 1999] (RBF, do original em inglês *radial basis function*) como κ.

2.2 Erro Quadrático Médio

Para verificarmos o desempenho do método de regressão proposto, utilizamos o erro quadrático médio (MSE, do original em inglês *Mean Squared Error*) que é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e^2 \tag{2.2}$$

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 seria a distância do valor de qualidade estimado pela regressão e o valor atribuído por estes avaliadores.

2.3 *NDCG@k*

Como a qualidade dos artigos pode ser tranformada em um *ranking* onde os artigos com a qualidade mais alta assumem as melhores posições, foi utilizado o Desconto Normalizado do Ganho Cumulativo nas k primeiras posições, também conhecido como

NDCG@k(do original em inglês *Normalized Discounted Cummulative Gain at top k positions*), para comparar o ranking produzido pelo método deste trabalho com os métodos propostos anteriormente na literatura.

Esta métrica, proposta por Järvelin & Kekäläinen [2000], demonstra a proximidade que um determinado *ranking* está do *ranking* que seria considerado o ideal.

O NDCG@k é definido da seguinte forma:

$$NDCG@k = \frac{1}{N}\sum_{i=1}^{k}(\frac{r_i}{\log_2(i+1)}) \tag{2.3}$$

onde r_i é a relevância do item que está na posição i do ranking e N é a norma calculada através do valor NDCG@k não normalizado ($N = 1$) de um ranking ideal onde $r_1 \geq r_2 \geq r_3 \geq ... \geq r_{k-1} \geq r_k$.

Ao estimarmos a qualidade de um artigo, a sua relevância seria a pontuação relativa à sua qualidade atribuída pelos avaliadores.

2.4 Ganho de Informação

O ganho de informação [Witten & Frank, 1999](ou info-gain do original em inglês *information gain*) é utilizado para avaliar o ganho de informação que cada atributo fornece ao tentar distinguir classes. Para calcular o ganho de informação, é necessário calcular o valor da informação, ou entropia, da seguinte forma:

$$informacao(p_1, p_2, p_3, ..., p_n) = entropia(\frac{p_1}{m}, \frac{p_2}{m}, \frac{p_3}{m}, ..., \frac{p_n}{m}) \tag{2.4}$$

$$entropia(q_1, q_2, ..., q_n) = q_1 \times log_2(q_1) - q_2 \times log_2(q_2) - ... - q_n \times log_2(q_n) \tag{2.5}$$

onde p_i representa o número de itens que possuem a classe i e m é o número total de itens. Quanto maior é a entropia, maior é o nível de desordem, ou seja, falta de informação.

Por exemplo, a Tabela 2.1 apresenta um conjunto de 7 artigos fictícios com suas classes e atributos discretizados. Desta forma, usando a equação 2.4, a informação da amostra como um todo é equivalente a $informacao(4, 3)$, pois existem 4 artigos bons e 3 ruins:

$$informacao(4, 3) = entropia(\frac{4}{7}, \frac{3}{7}) \tag{2.6}$$

$$entropia(\frac{4}{7}, \frac{3}{7}) = -\frac{4}{7} \times log_2\frac{4}{7} - \frac{3}{7} \times log_2\frac{3}{7} = 0,985 \tag{2.7}$$

Tabela 2.1. Exemplo de artigos/atributos discretizados e suas classes (estipuladas manualmente por avaliadores)

ID do artigo	Tamanho	Número de Seções	Número de Links	Classe
1	Grande	Poucas	Muitos	bom
2	Grande	Poucas	poucos	ruim
3	Médio	Muitas	Muitos	bom
4	Médio	Médio	Muitos	bom
5	Pequeno	Muitas	Muitos	ruim
6	Pequeno	Poucas	Muitos	ruim
7	Pequeno	Muitas	Muitos	bom

Logo após, para calcular o ganho de informação ao utilizar o atributo 'tamanho', por exemplo, deve-se calcular a informação quando o tamanho do artigo for grande, médio e pequeno. Sendo assim, o resultado seria $informacao(1,1) = 1$, $informacao(2,0) = 0$ e $informacao(2,1) = 0,918$, respectivamente. Em seguida, conforme Eq. 2.8, é calculado a média destes 3 resultados ponderada pela quantidade de artigos grandes, médios e pequenos. Finalmente, o info-gain é a diferença entre a informação antes de utilizar o atributo de tamanho (0.985) e a informação ao utilizar o atributo tamanho (Eq. 2.9).

$$info_{tamanho} = \frac{2}{7} \times 1 + \frac{2}{7} \times 0 + \frac{3}{7} \times 0,918 = 0,679 \tag{2.8}$$

$$infoGain(tamanho) = informacao(4,3) - info_{tamanho} \tag{2.9}$$

$$infoGain(tamanho) = 0,985 - 0,679 = 0,306$$

Em casos reais, como no presente trabalho, os atributos geralmente são numéricos. Sendo assim, é realizado a discretização dos atributos antes de calcular o info-gain, conforme demonstrado em Witten & Frank [1999].

Capítulo 3

Modelagem do Problema

3.1 Utilização do SVR para estimar a qualidade de artigos

Como a qualidade de um artigo pode ser vista como uma escala, onde um artigo pode estar mais distante ou mais próximo de uma classe, neste trabalho foi utilizado SVR, que, conforme já mencionado, é uma técnica de regressão que utiliza aprendizado de máquina. Seja $A = \{a_1, a_2, ..., a_n\}$ um conjunto de artigos, onde cada artigo a_i é representado por um conjunto de m atributos $\{F_1, F_2, ..., F_m\}$. Assim, $a_i = (f_{i1}, f_{i2}, ..., f_{im})$, onde f_{ij} é o valor do atributo F_j no artigo a_i.

Assumiu-se nesta proposta que se possui o acesso a um *conjunto de treino* no formato $\{(a_1, q_1), (a_2, q_2), ..., (a_n, q_n)\} \subset A \times \mathbb{R}$, onde cada par (a_i, q_i) representa um artigo a_i e o valor da avaliação correspondente a sua qualidade q_i, tal que se $q_1 > q_2$, então a qualidade do artigo a_1 é maior do que o artigo a_2, segundo o avaliador humano. Sendo assim q_i seria o 'resultado alvo' definido na Seção 2.1.

A solução proposta para este problema consiste em: (1) determinar o conjunto de atributos $\{F_1, F_2, ..., F_m\}$ utilizados para representar os artigos em A e (2) utilizar o SVR para encontrar a melhor combinação de atributos para estimar a qualidade q_i de um artigo a_i. Na próxima seção, serão apresentados os atributos utilizados para representar os artigos.

3.2 Representação do Artigo

Para estimar a qualidade de artigos utilizando regressão é essencial determinar quais atributos serão utilizados. Neste trabalho, foram utilizados atributos propostos an-

teriormente na literatura bem como outros novos, sugeridos aqui. Tais atributos são baseados nos critérios utilizados pela Wikipédia [Wikipedia, 2008e], 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.

Nas próximas seções serão apresentados os 68 atributos adotados no presente trabalho. Note que procuramos usar atributos que não fossem muito específicos da Wikipédia para tornar nossos resultados mais facilmente aplicáveis a coleções criadas por outras comunidades colaborativas.

3.2.1 Atributos do Histórico de Revisões

Os atributos extraídos do histórico de revisão são úteis para estimar a maturidade e estabilidade do texto [P. Dondio & Weber, 2006]. Espera-se que um artigo atinja um nível de maturidade suficiente para não necessitar de grandes alterações tornando-se assim, estável. Além disso, a falta de estabilidade no texto é característica de um conteúdo controverso ou que não reflete um ponto de vista neutro sobre o assunto, aspectos que devem ser evitados em textos de uma enciclopédia. Os atributos estudados são:

- **Idade (*hida*)**: Idade (em dias) do artigo. Um artigo de boa qualidade geralmente não é um artigo recente, pois não houve tempo suficiente para que muitas pessoas o revisassem.

- **Número de revisões (*hnr*)**: Número total de revisões. A intuição aqui é que quanto mais revisado o artigo, melhor o seu conteúdo.

- **Número de Revisões por IPs e usuários (*hni*,*hnu*)**: Número de revisões realizadas por usuários anônimos(hni) e registrados(hnu). Usuários anônimos, geralmente, não são tão comprometidos com o projeto quanto os registrados.

- **Média de Revisões por usuário (*hmru*)**: Definido na eq. 3.1, este atributo é útil para verificar a distribuição das revisões feitas pelos usuários.

$$hmru = \frac{hnr}{numUsuarios} \tag{3.1}$$

  onde $numUsuarios$ é o número de usuários que revisaram o artigo.

- **Desvio padrão das revisões por usuários (*hdpru*)**: Verifica o balanceamento das revisões entre os autores.

- **Número de revisões na página de discussão (*hnd*)**: Este atributo é útil para verificar a quantidade de discussões que são feitas em um artigo. Geralmente, alterações maiores devem ser discutidas em grupo para diminuir a chance do autor expressar seu ponto de vista.

- **Porcentagem de linhas modificadas (*hplm*)**: Número de linhas modificadas quando comparou-se a versão atual do artigo com a versão de 3 meses atrás. Este é um bom indicador da estabilidade de um artigo.

- **Porcentagem de revisões de usuários ocasionais (*hpuo*)**: Porcentagem de revisões realizadas por usuários que editaram o artigo menos que 4 vezes.

- **Porcentagem de revisão nos últimos 3 meses (*hputm*)**:

$$hputm = \frac{numRevTresMeses}{hnr} \tag{3.2}$$

onde $numRevTresMeses$ é o número de revisões realizadas nos últimos 3 meses no artigo.

- **Porcentagem de revisão dos usuários mais ativos (*hprua*)**: Definido na eq. 3.3, é a porcentagem de revisões dos *top* 5% usuários mais ativos.

$$hprua = \frac{numRevisoesTop5}{hnr} \tag{3.3}$$

onde $numRevisoesTop5$ é o número de revisões dos top 5% usuários mais ativos.

- **ProbReview (*hprv*)**: Proposto por Hu et al. [2007] esta métrica tenta estimar a qualidade dos artigos da Wikipédia baseado na qualidade do artigo e de seus usuários revisores. Recursivamente, a qualidade dos revisores é baseada na qualidade dos artigos que ele revisou. Formalmente, a qualidade de um artigo i, Q_i, é calculada da seguinte forma:

$$Q_i = \sum_k q_{ik} \tag{3.4}$$

onde q_{ik} é a qualidade do *k-ésimo* termo do artigo i. A qualidade de um artigo é definida como a probabilidade do usuario ter modificado cada termo do artigo multiplicado pela qualidade do usuário, como exposto na Eq. (3.5).

$$q_{ik} = \sum_j prob(w_{ik}, u_j)U_j \tag{3.5}$$

A qualidade do usuário é obtida pelo somatório da probabilidade do usuário ter revisado cada palavra de cada revisão que ele realizou, multiplicado pela qualidade de cada palavra, conforme Eq. (3.6):

$$U_j = \sum_{i,k} prob(w_{ik}, u_j) q_{ik} \tag{3.6}$$

A probabilidade $prob(p, u)$, definida na Eq. 3.7, do usuário u revisar o termo p é igual a 1 se u criou a palavra p. Caso contrário, a probabilidade é reduzida gradativamente de acordo com a distância entre p e o termo mais próximo criado por u, usando a formula:

$$prob(p, u) = \frac{1}{\sqrt{max(|d_{pl}| - 7, 0) + 1}} \tag{3.7}$$

onde d_{pl} é a distância entre o termo p e o termo l. Sendo que l é o termo mais próximo de p também criado pelo usuário u.

O indício *hida* foi proposto por Rassbach et al. [2007]; Mingus [2008] os demais atributos foram propostos por P. Dondio & Weber [2006]. Os seguintes atributos foram propostos pelo presente trabalho:

- **Idade por revisão (*hipr*)**: Utilizado para verificar a quantidade de tempo médio que um artigo passa sem ser revisado nas últimas 30 revisões. Este valor é obtido através da média de dias que o artigo passa sem ser revisado entre as últimas 30 revisões, excluindo o número máximo e mínimo de dias.
- **Revisões por dia (*hrpd*)**: Definido na eq. 3.8, este atributo verifica a frequência que um artigo é revisado.

$$hrpd = \frac{hnr}{hida} \tag{3.8}$$

Quanto ao custo computacional para o pré-processamento destes atributos, ProbReview (*hprv*) possui um custo quadrático de acordo com o número de revisões. A idade (*hida*) possui custo linear levando em conta o números de artigos e, as demais, possuem o custo linear levando em conta o número de revisões.

3.2.2 Atributos de Redes Complexas

Atributos de redes complexas são extraídos através do grafo de artigos existente na coleção. Segundo Barroso [1998] "grafo é uma estrutura matemática constituída de

dois conjuntos, um conjunto V, finito e não vazio, de n vértices, e outro E, de m arestas, que são pares não ordenados de elementos de V". No presente trabalho, os vértices são artigos da Wikipédia e as arestas são os apontadores entre um artigo e outro.

Atributos de redes complexas são relevantes para fornecer evidência da importância e popularidade de um artigo. Um dos motivos de se analisar a importância e popularidade dos artigos é identificar o motivo da estabilidade de um artigo, pois, segundo P. Dondio & Weber [2006], um artigo pode ser estável não só devido a sua maturidade, mas por causa da falta de importância do tópico ou sua baixa popularidade. Os atributos de redes complexas utilizados são:

- **PageRank (*rpr*)**: Utilizado em Rassbach et al. [2007], é calculado como definido em [Brin & Page, 1998]. Ele se baseia na idéia de que a importância de um artigo é proporcional à importância dos artigos que apontam pra ele, sendo estes ponderados por sua própria importância. Este atributo tenta estimar a popularidade de um artigo. Para calculá-lo, utilizou-se o programa *Web Graph* [Boldi & Vigna, 2004].

- **Grau de entrada da aresta (*rge*)**: Número de vezes que um artigo é referenciado por outros artigos. Este atributo fornece outro meio de se calcular a popularidade de um artigo com custo razoavelmente baixo.

- **Grau de saída da aresta(*rgs*)**: Utilizado em Rassbach et al. [2007]; Mingus [2008]; P. Dondio & Weber [2006] é o número de vezes que o artigo referencia outros artigos da Wikipédia. Um bom artigo deve possuir referências à outros artigos da Wikipédia.

- **Número de apontadores (*rnl*)**: Número de apontadores do artigo, utilizado em Rassbach et al. [2007]; Mingus [2008]; P. Dondio & Weber [2006]. Esta métrica é diferente do grau de saída(*rgs*) uma vez que o número de apontadores leva em conta também os apontadores para artigos que ainda não foram escritos, ou seja, apontadores quebrados.

- **Número de traduções (*rnl*)**: Proposto por Rassbach et al. [2007]; Mingus [2008] esta métrica calcula o número línguas em que este artigo já foi traduzido e disponibilizado na Wikipédia. Este atributo indica a importância, popularidade e universalidade do tema discutido pelo artigo.

Os seguintes atributos foram utilizados anteriormente na literatura para detectar spam em páginas na *Web* e através de sistemas de compartilhamento de vídeos online,

como o YouTube[1]. Porém, antes do presente trabalho, eles não haviam sido utilizados para estimar a qualidade de artigos:

- **Assortatividade entrada-entrada, entrada-saída, saída-entrada e saída-saída (*raee*, *raes*,*rase*,*rass*)**: Utilizado em Benevenuto et al. [2008]; Castillo et al. [2007], Assortatividade é definido nas Eqs. 3.9 , 3.10 , 3.11 , 3.12

$$raee = \frac{rge}{avgGrauEntrada} \tag{3.9}$$

$$raes = \frac{rge}{avgGrauSaida} \tag{3.10}$$

$$rase = \frac{rgs}{avgGrauEntrada} \tag{3.11}$$

$$rass = \frac{rgs}{avgGrauSaida} \tag{3.12}$$

  onde $avgGrauEntrada$ é a média do grau de entrada(rge) dos artigos vizinhos e $avgGrauSaida$ é a média do grau de saída(rgs) dos artigos vizinhos. Esta métrica tenta estimar semelhanças entre a aresta atual e seus vizinhos. Por exemplo, se o artigo atual possuir o grau de entrada alto e uma assortatividade entrada-entrada igual a um, pode significar que este artigo é tão popular quanto seus vizinhos.

- **Coeficiente de clusterização (*rcc*)**:

$$rcc = \frac{numArestas(k)}{maxArestas(k)} \tag{3.13}$$

  Onde $numArestas(k)$ representa o número de arestas existentes a partir do vértice atual até uma distância k deste vértice e $maxArestas(k)$ é o número máximo de arestas que podem existir até a distância k deste vértice. Utilizado em Benevenuto et al. [2008]; Dorogovtsev & Mendes [2003], este atributo indica se um artigo pertence a um grupo de artigos que se relacionam entre si. Um artigo que pertence a um grupo geralmente é mais acessível ao usuário.

- **Reciprocidade (*rre*)**: Porcentagem de artigos citados pelo artigo atual que também citam o artigo atual. Esta métrica tenta verificar a qualidade dos apontadores entre um artigo e outro. Pois, se a reciprocidade é alta, pode-se supor que o artigo atual geralmente aponta para artigos de assuntos relacionados.

[1] http://www.youtube.com

3.2.3 Atributos do Texto

Esses são atributos extraídos do conteúdo textual dos artigos. Como metade dos atributos apresentados neste trabalho são textuais, estes foram divididos em 4 sub-grupos: tamanho, estilo, estrutura e legibilidade.

A vantagem da utilização de atributos de texto é o menor custo de pré-processamento quando comparados com atributos que levam em conta o histórico de revisões e análise de apontadores. O custo de pré-processar atributos de texto é linear quanto ao tamanho dos artigos da amostra.

Além disso, atributos de texto são mais simples de extrair em qualquer outra enciclopédia colaborativa, facilitando assim a utilização destes atributos em outras coleções.

3.2.3.1 Atributos Relacionados ao Tamanho

Estes atributos, usados por Hu et al. [2007]; P. Dondio & Weber [2006]; Rassbach et al. [2007], são indicadores do tamanho do artigo. A intuição que está por trás destes atributos é que artigos completos e maduros não são artigos pequenos e incompletos nem grandes e prolixos.

- **Número de Caracteres (ttnc):** Número de caracteres no texto, incluindo espaços.
- **Número de frases e palavras (*ttnf*,*tlnp*):** Estes dois atributos correspondem ao número de palavras e frases contidas no texto, repectivamente.

3.2.3.2 Atributos de Estrutura

Atributos de estrutura fornecem indícios da organização do texto. Segundo os padrões de qualidade da Wikipedia [2008e], um bom artigo deve ser organizado em seções, fornecendo também referências e apontadores para fontes externas à Wikipédia. Sendo assim, estes atributos foram utilizados para tentar estimar a organização através dos apontadores, seções, imagens e citações existentes no texto. Os atributos utilizados neste trabalho foram:

- **Número de Seções (*tens*):** Número de seções do artigo. A intuição por trás desta atributo é que um artigo de qualidade é organizado em seções. Especificamente, na Wikipédia, um artigo completo geralmente possui no mínimo uma seção de introdução, uma de desenvolvimento e lista de referências e apontadores externos.

- **Desvio padrão do tamanho da seção (*tedpts*)**: Representa o balanceamento e distribuição das seções no texto. A intuição por trás deste atributo é que o conteúdo de um bom artigo não está concentrado em poucas seções.
- **Número de subseções (*tensb*)**: A intuição por trás destes atributos é que grandes seções de um texto geralmente são sub-divididas em seções menores.
- **Número de citações (*tenc*)**: Número de citações de um artigo. Um artigo de boa qualidade deve referenciar as suas fontes.
- **Número de ligações externas (*tenle*)**: Número de referências externas à Wikipédia no texto. Um bom artigo deve possuir informações adicionais que estejam disponíveis online na *Web*.
- **Número de figuras (*tenf*)**: Número de figuras no artigo. Figuras podem contribuir para que o conteúdo fique mais claro e visualmente mais adequado.

Os atributos *tens*, *tedpts*, *tenc*, *tenle*e *tenf* foram propostos por P. Dondio & Weber [2006] para estimar a qualidade de edição e *tensb* foi proposto por Rassbach et al. [2007]; Mingus [2008]. Os seguintes atributos foram propostos pelo presente trabalho:

- **Média do tamanho das seções (*temts*)**:

$$temts = \frac{ttnc}{tens} \tag{3.14}$$

- **Tamanho médio do parágrafo (*tetmp*)**:

$$tetmp = \frac{ttnc}{numParagrafos} \tag{3.15}$$

onde $numParagrafos$ é o número de parágrafos do texto.

- **Tamanho da maior e menor seção (*tetms*,*tetns*)**: Estes atributos são úteis para detectar alguma forma não usual de seções, ou muito grandes (que poderia ser dividido em sub-seções), ou muito pequenas (que poderia indicar um texto incompleto).
- **Média de subseções por seção (*temsbs*)**: Esta atributo indica a organização das seções e a distribuição de subseções ao longo das seções.

$$temsbs = \frac{tensb}{tens} \tag{3.16}$$

- **Tamanho do resumo (*tetr*)**: O tamanho em número de caracteres do texto introdutório ao assunto. Artigos mais maduros tendem a possuir um texto, antes do ínicio das seções, que resume o seu conteúdo.

- **Número de citações pelo tamanho do texto (*tsctcl*)**: Esta é uma versão relativa do *tenc*, que leva em conta o tamanho do artigo, pois artigos maiores possuem mais informações, portanto necessitam de mais referências.

$$tenct = \frac{tenc}{ttnc} \tag{3.17}$$

- **Números de citações por seção (*tencs*)**: Outra versão relativa da atributo *tenc* que leva em conta o tamanho do texto, porém desta vez verificando o número de seções. Um bom texto deve possuir citações na maior parte de suas seções.

$$tencs = \frac{tenc}{tens} \tag{3.18}$$

- **Apontadores pelo tamanho do texto (*teltt*)**: Através deste atributo é verificado a distribuição dos apontadores no texto. É calculado da seguinte forma:

$$teltt = \frac{rnl}{ttnc} \tag{3.19}$$

- **Números de ligações externas por seção (*tenles*)**: Este atributo foi criado para normalizar o número de apontadores externos. Geralmente, um texto com mais seções deve possuir mais apontadores externos para cobrir os assuntos das seções de forma adequada.

$$tenles = \frac{tenle}{tens} \tag{3.20}$$

- **Figuras por seção (*tenfs*)**: Este atributo indica o número de imagens no texto por seção.

$$tenfs = \frac{tenf}{tens} \tag{3.21}$$

3.2.3.3 Atributos de Estilo

Estes atributos têm a intenção de identificar a forma característica que os autores escrevem através da análise da distribuição de palavras e frases usadas no texto. A intuição por trás destes atributos é que artigos de boa qualidade possuem formas características de usar frases e palavras como, por exemplo, empregando frases pequenas. Estes atributos foram propostos por Rassbach et al. [2007] e, para calculá-los, utilizou-se o

programa Style and Diction [Style & Diction, 2008]. Na Tabela 3.1 são apresentados os termos utilizados para calcular cada atributo.

Tabela 3.1. Termos utilizados no cálculo de cada atributo de estilo

Atributo	Termos utilizados
tsnva	*will, shall, cannot, may, need to, would, should, could, might, must, ought, ought to, can't, can*
tsnpr	*I, me, we, us, you, he, him, she, her, it, they, them, thou, thee, ye, myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself, my, mine, his, hers, yours, ours, theirs, its, our, that, their, these, this, those, your*
tspc,*tsic*	*and, but, or, yet, nor*
tsps(sufixos)	*tion, ment, ence, ance*
tspp,*tsipr*	*aboard, about, above, according to, across from, after, against, alongside, alongside of, along with, amid, among, apart from, around, aside from, at, away from, back of, because of, before, behind, below, beneath, beside, besides, between, beyond, but, by means of, concerning, considering, despite, down, down from, during, except, except for, excepting for, from among, from between, from under, in addition to, in behalf of, in front of, in place of, in regard to, inside of, inside, in spite of, instead of, into, like, near to, off, on account of, on behalf of, onto, on top of, on, opposite, out of, out, outside, outside of, over to, over, owing to, past, prior to, regarding, round about, round, since, subsequent to, together, with, throughout, through, till, toward, under, underneath, until, unto, up, up to, upon, with, within, without, across, along, by, of, in, to, near, of, from*
tsvtb	*be, being, was, were, been, are, is*
tsia	*the, a, an*
tsics	*after, because, lest, till, 'til, although, before, now that, unless, as, even if, provided that, provided, until, as if, even though, since, as long as, so that, whenever, as much as, if, than, as soon as, inasmuch, in order that, though, while*
tsipi	*why, who, what, whom, when, where, how*

- **Tamanho da maior frase (tstmf)**: Número de palavras da maior frase.
- **Porcentagem de frases grandes (tspfg)**: Porcentagem de frases com 10 palavras a mais que a média do número total de palavras no texto.
- **Porcentagem de frases pequenas (tspfp)**: Porcentagem de frases que possuem 5 palavras a menos que a média total de palavras no texto.
- **Número de verbos auxiliares, perguntas, pronomes e frases na voz passiva (*tsnva*, *tsnp*, *tsnpr*, *tsnfp***: Estes atributos capturam o número de verbos, perguntas, pronomes e frases na passiva existentes no texto.

- **Porcentagem de conjunções (*tspc*)**:

$$tspc = \frac{cjCount}{tlnp} \tag{3.22}$$

onde $cjCount$ é o número de conjunções existentes no texto. Segundo [Faraco & de Moura, 1997] conjunção é uma palavra invariável que liga duas orações.

- **Porcentagem de substantivação (*tsps*)**: Substantivação, segundo Faraco & de Moura [1997], é a atribuição de função de substantivo a alguma outra palavra.

$$tsps = \frac{sbCount}{tlnp} \tag{3.23}$$

onde $sbCount$ é o número de palavras que foram transformadas em substantivos no texto.

- **Porcentagem de preposições (*tspp*)**: Porcentagem de preposições no texto.

$$tspp = \frac{prepCount}{tlnp} \tag{3.24}$$

onde $prepCount$ é o número de preposições no texto.

- **Porcentagem de verbos *'to be'* (*tsvtb*)**: Porcentagem do verbo auxiliar *to be* no texto.

$$tsvtb = \frac{toBeCount}{tlnp} \tag{3.25}$$

onde $toBeCount$ é o número de verbos to-be no texto.

- **Atributos relacionados ao início de frases**: Número de frases que iniciam com pronome($tsip$), artigo($tsia$), conjunções($tsic$), conjunções subordinativas($tsics$), pronomes interrogativos ($tsipi$) e preposição($tsipr$).

3.2.3.4 Atributos de Legibilidade

Atributos de legibilidade, utilizados por Rassbach et al. [2007], pretendem estimar a idade ou o nível de instrução (estimado de acordo com o nível de escolaridade nos EUA) que a pessoa deve possuir para conseguir compreender um texto. Estes atributos utilizam a combinação de vários indícios como número e tamanho de palavras, frases e sílabas com o objetivo de verificar o grau de complexidade do texto. Como os artigos da Wikipédia são dirigidos para um público não especializado no assunto, a intuição

por trás destes atributos é que bons artigos devem ser bem escritos, compreensíveis e livres de quaisquer complexidades desnecessárias. Os atributos utilizados são:

- **Índice Automatizado de Legibilidade (tlari)**: Esta métrica, do original inglês *Automated Readability Index*, foi proposta por Smith & Senter [1967] e consiste na utilização do número de palavras por frase e letras por palavras para estimar a legibilidade do texto:

$$ARI = 4.71\frac{numCaracteres}{numPalavras} + 0.5\frac{numPalavras}{numFrases} - 21.43 \tag{3.26}$$

- **Coleman-Liau (*tlcl*)**: Esta métrica foi proposta por Coleman & Liau [1975] e utiliza o número de letras por palavras e o número de frases num fragmento de 100 palavras ($wfCem$).

$$COLEMAN = 5.89\frac{numCaracteres}{numPalavras} - 0.3wfCem - 15.48 \tag{3.27}$$

- **Flesch *reading ease* (tlfe) e Flesch-Kincaid (*tlfk*)**: *tlfe* foi proposto por Flesch [1948] e *tlfe* foi demonstrado por Ressler [1993]. Estas métricas utilizam a média do tamanho das frases e de sílabas para estimar a legibilidade do texto. A primeira métrica retorna um valor de 0 a 100, onde 0 indica um texto difícil de entender.

$$FleshEase = 206.835 - 1.015\frac{numPalavras}{numFrases} - 84.6\frac{sílabas}{numPalavras} \tag{3.28}$$

  A segunda retorna o nível de instrução ao invés do valor entre 0 e 100.

$$FleshKincaid = 0.39\frac{numPalavras}{numFrases} + 11.8\frac{sílabas}{numPalavras} - 15.59 \tag{3.29}$$

- ***Gunning Fog Index* (tlgf)**: Esta métrica proposta em Gunning [1952] utiliza a média de palavras por frase e a média de palavras complexas utilizadas no texto.

$$GunningFog = 0.4(\frac{numPalavras}{numFrases} + 100\frac{palavrasComplexas}{numPalavras}) \tag{3.30}$$

  onde *palavrasComplexas* é definida como palavras que possuem três ou mais sílabas.

- **Läsbarhetsindex (tllix)**: Esta métrica, definida por Björnsson [1968], é utilizada para estimar a legibilidade do texto em várias línguas. É semelhante ao

tlgf, porém em *trlix* palavras complexas são definidas como palavras acima de 6 letras. O resultado é um valor positivo, sendo que o valor retornado aumenta à medida que a compreensão do texto se torna mais difícil.

$$Lix = \frac{numPalavras}{numFrases} + 100\frac{palavrasComplexas}{numPalavras} \tag{3.31}$$

- ***Smog-Grading* (*tlsg*)**: Esta métrica foi proposta por McLaughlin [1969] e utiliza o número de palavras polissílabas no texto. Esta métrica geralmente é utilizada em documentos relacionados à área da saúde.

$$Smog = 3 + \sqrt{polissilabas} \tag{3.32}$$

onde *polissilabas* é o número de palavras polissílabas (excluindo nomes próprios) extraídas de 30 frases do texto.

Para calcular estes atributos, também foi utilizado o programa *Style and Diction* [Style & Diction, 2008].

Capítulo 4

Experimentos

Utilizando os atributos da Seção 3.2 foram realizados uma série de experimentos. Neste capítulo será apresentado a metodologia adotada nos experimentos e seus resultados.

4.1 Coleção Utilizada

Foi utilizada uma amostra da Wikipédia, para os experimentos. Esta amostra foi extraída do repositório global, em inglês, contendo aproximadamente 2.683.000 artigos [Wikipedia, 2008a]. Deste repositório, aproximadamente 900.000 artigos já foram classificados quanto à sua qualidade, manualmente [Wikipedia, 2008d].

Os artigos da Wikipédia são de livre escrita e leitura, com exceção dos vandalizados. Estes podem ser editados apenas por usuários registrados. Além disso, os artigos possuem uma página de discussão, onde usuários podem discutir suas edições, resolver conflitos, criticar e avaliar o artigo. A avaliação do artigo pode ser feita por qualquer usuário da Wikipédia, obedecendo a seguinte escala de qualidade[1] [Wikipedia, 2008d]:

- *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.

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

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

- *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.

Qualquer pessoa pode classificar um artigo na Wikipédia, exceto na classificação de *FA* e *GA* onde, primeiramente, algum usuário registrado deve nomeá-lo. Após a nomeação, o artigo passa por uma fase de votação, realizada também por usuários registrados, para verificar se este artigo atende aos requisitos da classe [Wikipedia, 2008b,c].

Os artigos da Wikipédia geralmente são avaliados por grupos de usuários que fazem parte de um projeto. Um artigo pode múltiplas qualificações, já que este pode ser classificado por vários grupos e cada um com uma opinião distinta. Além disso, a qualidade, principalmente em classes vizinhas, pode ser bastante subjetiva. Portanto, para não lidar com o problema de multi-classificação, usamos somente artigos avaliados em apenas uma classe.

Os seguintes fatores foram importantes para determinar o tamanho da amostra. Alguns atributos, como *ProbReview* e de *Redes Complexas*, possuem um custo computacional mais alto para serem calculados. Além disso, para calcular o *ProbReview* é necessário fazer o *download* de todas as revisões de cada artigo. Para estudar o impacto e implementar atributos novos e propostos anteriormente na literatura, uma série de experimentos foram realizados, o que tornaria difícil lidar com uma amostra muito grande. Desta forma, foi coletado uma amostra aleatória e balanceada com aproximadamente 920 artigos. Após remover artigos com mais de uma classe, artigos de redirecionamento e artigos que na verdade eram listas, esta amostra ficou com 874 artigos e ligeiramente desbalanceada. Estes 874 artigos, coletados em janeiro de 2008, possuem 806.545 revisões, totalizando 38Gb de texto.

Na Tabela 4.1 é apresentada a distribuição de classes da amostra coletada e a distribuição de classes real. A informação sobre a classe do artigo foi extraída através de sua página de discussão. Nesta tabela também é possível observar que a classe que obteve menos artigos na amostra foi justamente a classe que possuia menos artigos na coleção completa.

Os atributos de redes complexas foram obtidos através dos apontadores entre as páginas da Wikipédia. Estes apontadores foram extraídos através de um arquivo

Tabela 4.1. Distribuição dos artigos por classe na amostra e o total de artigos classificados na Wikipédia em inglês

Classe	FA	AC	GA	BC	ST	SB
# artigos na amostra	142	76	109	168	185	168
# total de artigos	1.868	553	2.935	47.847	272.571	722.002

de importação disponível para download[2]. O grafo completo obtido através destes apontadores possui 3.165.998 vértices (artigos e páginas que redirecionam para artigos) e 86.077.675 arestas (apontadores de uma pagina para outra). Foi utilizado o programa denominado *Web Graph* [Boldi & Vigna, 2004] para a compactação do grafo e cálculo dos atributos de redes complexas.

4.2 Metodologia de Avaliação

Os experimentos realizados tiveram o objetivo de analisar o impacto causado por cada grupo de atributos e comparar o método proposto no presente trabalho com outros anteriormente propostos na literatura.

Com o objetivo de auxiliar a avaliação do impacto dos atributos, foi utilizado o info-gain, definido na Seção 2.4. Esta métrica geralmente é utilizada para seleção de atributos, porém, como ela fornece, através do valor de info-gain, um ranking de atributos baseado no seu poder discriminativo, também é possível usá-la para avaliar o impacto dos atributos.

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*) a 5 (*Featured Article*). Sendo assim, com o objetivo de avaliar o desempenho deste método, utilizou-se o erro quadrático médio (MSE, do original em inglês *Mean Squared Error*) definido na Seção 2.2. No presente trabalho, o erro é a diferença entre o valor de qualidade estimado e o seu valor real (definido manualmente pelos avaliadores).

Cada método apresentado neste trabalho retorna um valor referente à qualidade do artigo. Para tanto, foi construído um *ranking* ordenando os artigos através deste valor. Logo após, utilizou-se o NDCG@k, definido na Seção 2.3, para comparar o *ranking* dos artigos do método atual com os propostos anteriormente na literatura. Neste trabalho, o valor de relevância do artigo r_i, definido na Seção 2.3, é o valor de qualidade real, definido pelos avaliadores humanos. A norma N, definida também

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

na Seção 2.3, foi calculada através do *ranking* dos artigos ordenados pelo seu valor de qualidade real. O método proposto por Dondio et al. [2006]; P. Dondio & Weber [2006] avaliou os resultados obtidos através de análise visual e Hu et al. [2007] usaram a métrica NDCG@k para comparar seus resultados com outros ao estimar a qualidade de artigos da Wikipédia.

Uma limitação que tivemos ao comparar nosso trabalho com os outros propostos na literatura foram as descrições vagas. Em geral, nem sempre indícios e métodos são descritos de maneira clara. Este foi o caso de grande parte dos indícios propostos por Rassbach et al. [2007]. Para compensar tal problema, as versões daqueles indícios apresentados foram obtidos através de comunicação pessoal com o autor. Também, ao implementar P. Dondio & Weber [2006], que cria rankings combinando indícios, estes indícios foram combinados de duas formas: utilizando a metodologia proposta no presente trabalho e por meio da média simples dos valores de seus indícios.

Nos métodos de aprendizado de máquina apresentados neste trabalho, os experimentos foram realizados de acordo com a metodologia de validação cruzada de dez partições (10-fold cross validation) [Mitchell, 1997]. 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 ($\frac{1}{10}$), 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.

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, 95% de confiança.

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

4.3 Resultados

4.3.1 Análise dos Atributos

Nesta seção é avaliado o impacto de cada atributo ao estimar a qualidade de artigos utilizando *SVR*. Devido ao grande número de atributos, torna-se impraticável uma análise considerando todas as suas combinações, pois o número de combinações possí-

veis cresce exponencialmente com o número de atributos. Sendo assim, agrupou-se os atributos em histórico de revisões, redes complexas e os grupos relacionados ao texto: tamanho, estilo, estrutura e legibilidade. Logo após, foram realizados dois estudos, um utilizando os grupos de atributos e outro analisando, dentro de cada grupo, os seus atributos individualmente.

Com o objetivo de analisar os grupos de atributos, foram conduzidas duas séries de experimentos. Primeiramente, foi analisado o resultado de cada grupo, para verificar seu impacto individual. Logo após, para analisar o impacto da falta de cada grupo, foram combinados todos os grupos e removidos um por vez.

Na Tabela 4.2 são apresentados os resultados obtidos na coleção de testes para cada grupo, quando utilizados individualmente (coluna *Único Grupo*) e quando este grupo é excluído do conjunto de todos os outros grupos (coluna *Grupo Excluído*). Nesta tabela é exibido, para cada grupo, o MSE obtido e o ganho percentual do MSE quando comparado ao resultado da combinação de todos os grupos ($MSE = 0,82$). Além disso, valores de MSE marcados com '*' na Tabela 4.2 indicam diferenças estatisticamente significativas em relação aos obtidos quando utilizando todos os grupos.

Tabela 4.2. Desempenho da regressão ao utilizar apenas um grupo (*Único Grupo*) para representar o artigo e todos os atributos exceto um grupo específico (*Grupo excluído*). Valor MSE para comparação = 0.82 (utilizando todos os grupos)

Grupo	**Único Grupo**		**Grupo excluído**	
	erro (MSE)	**aumento**	**erro (MSE)**	**aumento**
Estrutura	0,98*	19%	0,91*	11%
Tamanho	1,18*	44%	0,82	0%
Estilo	1,23*	50%	0,83	1%
Histórico de Rev.	1,24*	51%	0,90*	10%
Redes Comp.	1,38*	68%	0,85	3%
Legibilidade	2,72*	231%	0,83*	1%

Observando a Tabela 4.2, é possível perceber que o grupo de *Estrutura*, quando sozinho, conseguiu obter o melhor desempenho entre todos os outros grupos. Enquanto isso, métricas de *legibilidade* obtiveram o pior desempenho. Isto sugere que indícios relacionados à organização do artigo como seções, imagens e citações são melhores para distinguir a sua qualidade. Portanto, é possível concluir que os indícios de *legibilidade* não conseguem distinguir, isoladamente, artigos de alta e baixa qualidade.

Os grupos de *tamanho* e *estilo*, mesmo obtendo resultados relativamente bons quando utilizados de forma isolada, demonstraram pouco (ou nenhum) impacto quando excluído do grupo contendo todos os atributos. Isto pode ser explicado, em parte, pelo

fato de alguns atributos de *estilo* podem ser substituídos por atributos de *tamanho*, pois eles fornecem o mesmo tipo de informação. Por exemplo, *o número de frases que iniciam com preposições*, que é atributo de *estilo*, geralmente é proporcional ao tamanho do documento. Logo, a remoção dos atributos de *tamanho* e *estilo* não causam impacto no resultado final.

Também é possível observar, através de uma análise mais detalhada nos atributos de texto que utilizar apenas atributos de *estrutura* e *estilo* é tão bom quanto utilizar todos os atributos de texto combinados. Isto é observado na Tabela 4.3, onde é apresentado o MSE obtido por diversas combinações de atributos de texto. Além disso, atributos de *estrutura* + *estilo* conseguiram obter os melhores resultados, tanto isoladamente, quanto combinados com outros atributos.

Tabela 4.3. MSE de artigos representados por diversas combinações de atributos de texto.

Atributos	**erro (MSE)**
Estrutura + Estilo	0,92
Todos os atributos de texto	0,93
Estrutura + Estilo + Legibilidade	0,94
Estrutura + Estilo + Tamanho	0,94
Estrutura + Legibilidade	0,97
Estrutura + Tamanho	0,97
Estrutura + Legibilidade + Tamanho	0,98
Estilo + Legibilidade + Tamanho	1,10
Estilo + Tamanho	1,13
Legibilidade + Tamanho	1,16
Estilo + Legibilidade	1,19

Note que atributos de *legibilidade* não levam em consideração, direta ou indiretamente, se o artigo cobre de forma adequada o assunto abordado. Desta forma, até artigos rascunhos podem ser considerados de boa qualidade se possuírem uma boa legibilidade de acordo com as métricas da Seção 3.2.3.4. Isto explica o fraco desempenho dos atributos de *legibilidade* quando utilizados de maneira isolada (Tabela 4.2). Entretanto, quando combinado com outros atributos que consideram outros aspectos da qualidade do texto, os resultados melhoram consideravelmente (Tabela 4.3).

Na Tabela 4.2, é possível observar que os atributos de *Redes Complexas* geralmente possuem baixo impacto. Além disso, atributos do grupo de *Histórico de Revisões* contribuiram para a melhoria do resultados quando combinados com outros grupos.

Atributos do *Histórico de Revisões* são úteis para estimar a qualidade de artigos de boa e média qualidade, como pode ser observado na Figura 4.1. Nesta figura é exibido o desempenho da combinação dos atributos de *Estrutura* e *Estilo*, bem como

Estrutura, *Estilo* e *Histórico de Revisão*. O desempenho foi medido em termos de NDCG@k para valores crescentes de k. Pode-se observar um desempenho superior da combinação de apenas *Estrutura* e *Estilo*, quando $k < 150$. Porém, para valores entre 150 e 320, *Estrutura* e *Estilo* combinados com o *Histórico de Revisão* obtiveram um melhor desempenho. Lembrando que existem 142 artigos FA e 76 AC (Tabela 4.1), é possível concluir que o valor $k = 150$ corresponde, aproximadamente, à divisão entre artigos FA e AC. Estas duas classes são, de fato, mais difíceis de separar, pois artigos AC são candidatos a se tornar FA. Portanto, eles compartilham de características comuns. Além disso, AC possui um número reduzido de artigos na amostra, dificultando o processo de aprendizado. Isto pode explicar também a queda significativa de desempenho entre $k = 100$ e $k = 200$.

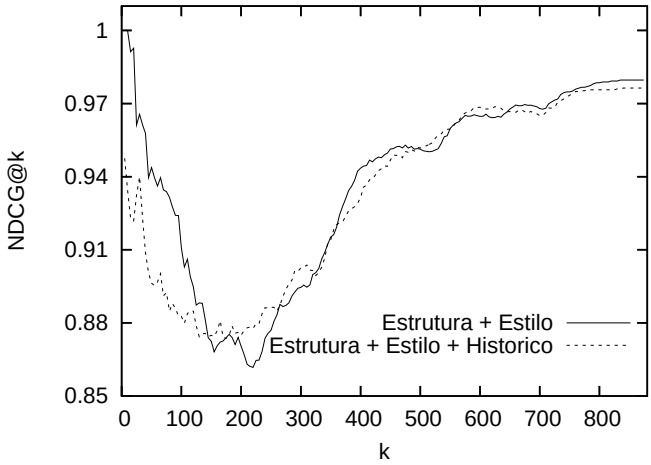

Figura 4.1. NDCG@k dos artigos, utilizando diferentes combinações de atributos.

O segundo experimento consiste em calcular o info-gain de cada atributo. Através deste valor é possível inferir a importância de um atributo na distinção de uma classe. Foram realizados experimentos mantendo apenas $N\%$ dos atributos de infogain mais alto, variando o valor de N e observando o impacto no resultado da regressão.

Verificando os atributos de *Estrutura*, podemos observar na Figura 4.2 que a maior perda neste grupo ocorreu quando foi utilizado apenas 40% dos atributos de mais alto infogain. Esta perda ocorreu quando foram retirados os atributos de *ligações externas* (*tenle*), *números de subseções* (*tensb*) e *tamanho do resumo* (*tetr*), o que confirma a importância destes atributos. As curvas dos atributos de *Estilo* e *Tamanho* não pareceram afetadas pela remoção de atributos, o que indica que estes atributos são similares entre si e, desta forma, capturam aspectos parecidos não agregando informação adicional para estimar a qualidade de um artigo.

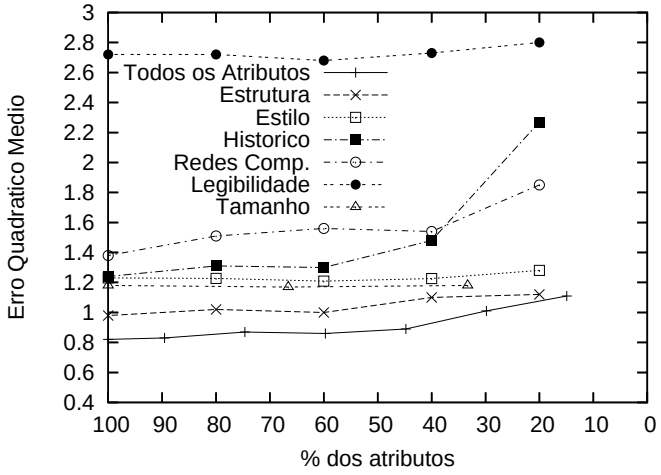

Figura 4.2. Erro quadrático médio obtido quando utilizado apenas $N\%$ dos atributos de maior info-gain.

Analisando os atributos de *Histórico de Revisão*, pode-se observar uma grande perda quando apenas 40% dos atributos foram utilizados. Esta perda ocorreu quando foram retirados os atributos *número de revisões por IPs* (*hni*), idade por revisão (*hipr*) e *Porcentagem de revisões de usuários ocasionais* (*hpuo*). Todos estes atributos ajudam a verificar a freqüência de revisão dos artigos e a qualidade dos usuários que realizaram estas revisões, já que usuários ocasionais e usuários não registrados estão menos envolvidos no processo de edição de um artigo. Assim, se um artigo é freqüentemente atualizado por estes usuários há uma maior probabilidade do mesmo possuir uma qualidade baixa.

Apesar do *ProbReview* (*hprv*) ser o atributo mais discriminativo do grupo de revisões conforme o Infogain, ele possui um custo computacional alto. Assim, para verificar o impacto deste atributo, foi realizado um experimento adicional retirando o *ProbReview* do grupo de *revisões*. O MSE dos atributos de revisão sem o *ProbReview* continuou o mesmo (1,24) após a retirada deste atributo. Pode se verificar também na Figura 4.2 que o *ProbReview*, mesmo combinado com os dois melhores atributos (20%), não consegue um bom desempenho. Sendo assim, conclui-se que o *ProbReview* pode ser descartado ao estimar a qualidade de artigos, devido o seu custo computacional e seu baixo impacto ao retirá-lo.

Verificando os atributos de *Redes Complexas*, pode-se observar que há uma grande perda de performance com apenas 20% dos atributos. Isto ocorre, não apenas pelo baixo número de atributos utilizados (*PageRank* (*rpr*) e *grau de saída* (*rgs*)), mas também pela importância dos dois atributos removidos: *coeficiente de clusterização* (*rcc*) e *número de apontadores* (*rnl*).

Tabela 4.4. Ranking dos atributos utilizando Info-gain

Ranking	**Info-gain**	**Atributo**	**Grupo**
1	0,9168	Número de caracteres (ttnc)	Tamanho
2	0,8973	Número de palavras (*tlnp*)	Tamanho
3	0,8531	Número de frases na passiva(*tsnfp*)	Estilo
4	0,8048	ProbReview (*hprv*)	Histórico de Revisões
5	0,8021	Número de frases (*ttnf*)	Tamanho
6	0,6963	Número de pronomes (*tsnpr*)	Estilo
7	0,6954	Número de revisões por usuários (*hnu*)	Histórico de Revisões
8	**0,6644**	**Tamanho da maior seção (*tetms*)**	**Estrutura**
9	0,6482	Número de revisões na página de discussão (*hnd*)	Histórico de Revisões
10	0,639	Número de frases que iniciam com artigo (*tsia*)	Estilo
11	0,6343	Número de revisões (*hnr*)	Histórico de Revisões
12	0,6233	Número de citações (*tenc*)	Estrutura
13	0,6135	Número de frases que iniciam com preposição (*tsipr*)	Estilo
14	0,6112	Número de seções (*tens*)	Estrutura
15	**0,6066**	**Número de revisões por dia (*hrpd*)**	**Histórico de Revisões**
16	0,592	Número de verbos auxiliares (*tsnva*)	Estilo
17	0,5724	Desvio padrão das revisões por usuário (*hdpru*)	Histórico de Revisões
18	**0,5695**	**Número de citações por seção (*tencs*)**	**Estrutura**
19	0,5271	Frases que iniciam com conjunção subordinativas	Estilo
20	**0,5078**	**Média do tamanho das seções (*temts*)**	**Estrutura**
21	0,4987	Frases que iniciam com pronome (*tsip*)	Estilo
22	0,4758	PageRank (*rpr*)	Redes Complexas
23	0,4688	Número de apontadores (*rnl*)	Redes Complexas
24	0,4668	Grau de saída da aresta (*rgs*)	Redes Complexas
25	**0,4445**	**Número de citações pelo tamanho do texto(*tenct*)**	**Estrutura**
26	**0,4358**	**Tamanho médio do parágrafo (*tetmp*)**	**Estrutura**
27	0,4255	Número de usuários ocasionais (*hpuo*)	Histórico de Revisões
28	0,4176	Número de revisões por IP (*hni*)	Histórico de Revisões
29	0,4091	Número de ligações externas (*tenle*)	Estrutura
30	**0,4022**	**Idade por revisão (*hipr*)**	**Histórico de Revisões**
31	0,3995	Média de Revisões por usuário (*hmru*)	Histórico de Revisões
32	**0,3942**	**Coeficiente de clusterização (*rcc*)**	**Redes Complexas**
33	**0,3423**	**Assortatividade saída-saída (*rass*)**	**Redes Complexas**
34	0,3392	Número de traduções (*rnl*)	Redes Complexas
35	0,3042	Idade (*hida*)	Histórico de Revisões
36	**0,2999**	**Tamanho do resumo (*tetr*)**	**Estrutura**
37	0,2872	Número de sub-seções (*tensb*)	Estrutura
38	0,2871	Grau de entrada da aresta (*rge*)	Redes Complexas
39	**0,2855**	**Apontadores pelo tamanho do texto (*teltt*)**	**Estrutura**
40	**0,2738**	**Média de subseções por seção (*temsbs*)**	**Estrutura**
41	0,2613	Tamanho da maior frase (*tstmf*)	Estilo
42	**0,2493**	**Assortatividade Entrada-Saída (*raes*)**	**Redes Complexas**
43	**0,2428**	**Assortatividade Saída-Entrada (*rase*)**	**Redes Complexas**
44	0,2416	Número de perguntas (*tsnp*)	Estilo
45	0,2397	Desvio padrão do tamanho da seção (*tedpts*)	Estrutura
46	0,2334	Frases que iniciam com pronome interrogativo (*tsipi*)	Estilo
47	**0,2043**	**Tamanho da menor seção (*tetns*)**	**Estrutura**
48	0,1945	Porcentagem de usuários ativos (*hprua*)	Histórico de Revisões
49	0,1669	Porcentagem de frases pequenas (*tspfp*)	Estilo
50	**0,1539**	**Assortatividade entrada-entrada (*raee*)**	**Redes Complexas**
51	0,1467	Porcentagem de revisão nos ultimos 3 meses (*hputm*)	Histórico de Revisões
52	0,1464	Porcentagem de frases grandes (*tspfg*)	Estilo
53	0,1436	Porcentagem de substantivação (*tsps*)	Estilo
54	0,1295	Porcentagem de conjunções (*tspc*)	Estilo
55	**0,1169**	**Figuras por seção (*tenfs*)**	**Estrutura**
56	**0,1091**	**Números de ligações externas por seção (*tenles*)**	**Estrutura**
57	0,1059	Porcentagem de verbos *'To be'* (*tsvtb*)	Estilo
58	0,0967	Coleman-Liau (*tlcl*)	Legibilidade
59	0,0891	Frases que iniciam com conjunções (*tsic*)	Estilo
60	0,0796	Läsbarhetsindex (*tllix*)	Legibilidade
61	**0,0766**	**Reciprocidade (*rre*)**	**Redes Complexas**
62	0,0763	Porcentagem de linhas modificadas (*hplm*)	Histórico de Revisões
63	0,0481	*Smog-Grading* (*tlsg*)	Legibilidade
64	0	Índice Automatizado de Legibilidade (*tlari*)	Legibilidade
65	0	*Gunning Fog Index* (*tlgf*)	Legibilidade
66	0	Flesch *reading ease*(*tlfe*)	Legibilidade
67	0	Flesch-Kincaid (*tlfk*)	Legibilidade
68	0	Porcentagem de preposições (*tspp*)	Estilo

Na Tabela 4.4, é exibido o ranking utilizando todos os atributos, ordenados pelo info-gain. Atributos em negrito foram propostos no presente trabalho. Esta tabela confirma a importância de atributos de texto para estimar a qualidade de artigos. Atributos deste grupo ocuparam 7 das 10 primeiras posições no *ranking*, sendo que as outras 3 foram ocupadas por atributos do Histórico de Revisões. Atributos relacionados ao tamanho do artigo ocuparam as 2 melhores posições, sendo que a quarta foi ocupada pelo atributo de revisão *ProbReview* que, segundo Hu et al. [2007], dada a definição de *ProbReview*, também incorpora informação do tamanho do texto. Entretanto, como se pode observar na Tabela 4.2 e 4.3, outros indícios que não levam em conta apenas o tamanho conseguem um melhor desempenho. Como o info-gain analisa cada atributo individualmente, pode-se concluir que o tamanho é o mais discriminativo quando analisado de forma individual, porém a combinação de atributos relacionados a outros aspectos do texto podem conseguir um melhor desempenho.

Nesta tabela também é possível verificar que os atributos de *Estrutura* obtiveram posições piores que o grupo de estilo e de tamanho no *ranking*. Mesmo assim, como é possível observar na Tabela 4.3, estes atributos conseguem um desempenho superior a estes dois outros grupos quando combinado com grupos de atributos de texto e mesmo quando avaliado de forma isolada (Tabela 4.2). Em outras palavras, os atributos de *Estrutura* conseguem resultados superiores quando combinados, mas não quando analisados de forma independente.

Atributos de revisão obtiveram boas colocações no *ranking*. Os mais bem colocados foram, em ordem, *ProbReview* (*hprv*), *Número de revisões por usuários* (*hnu*) e *Número de revisões na página de discussão* (*hnd*). Enfim, pode-se observar que os primeiros atributos de *Redes Complexas* aparecem apenas nos *top* 30 atributos de melhor resultado. Destes resultados, junto com o observado na Tabela 4.2 e o fato que atributos de *Redes Complexas* são computacionalmente caros, concluímos que o uso destes atributos pode ser evitado na estimativa da qualidade de artigos da Wikipédia. O mesmo pode-se concluir para atributos de legibilidade.

Na Tabela 4.5 temos a distribuição (em porcentagem) dos artigos com erros inferiores a um determinado valor, variando de 0.5 até 2.5. Como pode-se observar, artigos Stub-Class e Start-Class foram os mais facilmente classificados. Sendo que mais de 90% dos mesmos obtiveram um erro inferior a 1. Em geral, quanto mais alto é o nível de qualidade, maior é o erro ao estimá-lo. Artigos de classes baixas, podem ser claramente classificados através de algum indício de qualidade. Por exemplo, conforme mencionado na Seção 4.1, artigos da classe *Stub* são artigos de tamanho pequeno, enquanto *starter-class* são artigos pequenos, com alguma referência bibliográfica e estrutura. Esta distinção é progressivamente mais difícil para artigos de qualidade mais

alta. Além disso, observa-se que os piores resultados são com a classe *A-Class*, o que pode ser atribuído, conforme já mencionado, à dificuldade em distingui-la das classes FA e GA e ao menor número de artigos nesta classe na amostra utilizada.

Outro fator que contribui com o baixo desempenho dos artigos AC e um melhor desempenho das classes FA e GA é a forma diferenciada com que estas três classes são avaliadas manualmente. Conforme mencionado na Seção 4.1, artigos FA e GA possuem uma metodologia de avaliação mais rigorosa. Assim, a classificação manual destas duas classes de artigos é mais precisa, porém mais demorada. Por um lado, isto facilita o processo de treinamento com SVR dos artigos de classe GA e FA. Por outro lado, um artigo, provavelmente AC, pode possuir todas as características de um artigo FA, porém demorar mais para obter esta classe.

Classe	**% de artigos com erro ≤ a...**				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**
Featured-Article	29%	51%	74%	91%	97%
A-Class	21%	45%	64%	83%	97%
Good Article	39%	70%	86%	96%	100%
B-Class	46%	71%	88%	97%	99%
Start Class	74%	94%	99%	99%	100%
Stub	79%	96%	99%	100%	100%

Tabela 4.5. Distribuição do erro por classe

4.3.2 Comparação com Resultados Anteriores

Nesta seção comparou-se o metodo utilizado neste trabalho com outros métodos propostos na literatura. Especificamente, o presente método foi comparado com o método proposto por Dondio et al. [2006]; P. Dondio & Weber [2006] e Rassbach et al. [2007]. Estes métodos foram os que obtiveram os melhores resultados na literatura que, como o presente trabalho, utilizam combinações de indícios para estimar a qualidade de artigos da Wikipédia. Hu et al. [2007] também propôs uma combinação linear dos resultados obtidos com o atributo do histórico de revisão *ProbReview* e atributos relacionados ao tamanho do artigo, porém o resultado foi semelhante ao utilizar o *ProbReview* separadamente. Assim, este método não será avaliado neste trabalho.

A partir de agora, estes dois métodos serão chamados de DANTE e RASSBACH, respectivamente. Como os algoritmos descritos não foram disponibilizados publicamente, implementamos ambos os métodos. Para cada algoritmo, foi implementada duas versões. As primeiras versões dos algoritmos serão denominadas DANTE-LINEAR e RASSBACH-MAXENT, as quais correspondem aos métodos propostos originalmente.

Em particular, no caso de DANTE-LINEAR, foi feita a combinação das evidências necessárias para compor o ranking utilizando uma combinação linear. Em outras palavras, foi realizada a média de suas evidências.

Na segunda versão, que será denominada DANTE-SVR e RASSBACH-SVR, foram utilizados os atributos originalmente propostos pelos autores como entrada para o algoritmo SVR. Desta forma, é possível verificar a efetividade dos atributos propostos no presente trabalho quando comparados com os propostos por P. Dondio & Weber [2006] e Rassbach et al. [2007].

Como o algoritmo RASSBACH-MAXENT se trata de uma classificação, o resultado do mesmo foi transformado em um ranking utilizando a probabilidade de um artigo pertencer a uma determinada classe.

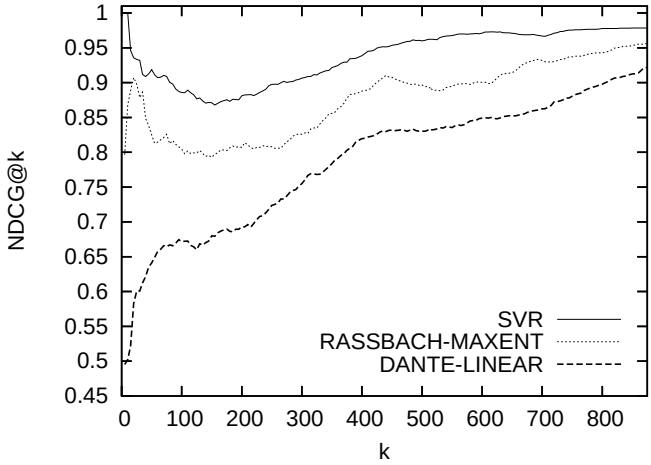

Figura 4.3. NDCG@k obtido pelos métodos DANTE-LINEAR, RASSBACH-MAXENT e SVR

Na Figura 4.3 é apresentado NDCG@k para DANTE-LINEAR, RASSBACH-MAXENT e a metodologia proposta neste trabalho, o SVR, utilizando o melhor conjunto de atributos para cada método. Nesta figura é possível verificar que os métodos baseados em aprendizado de máquina (RASSBACH-MAXENT e SVR) foram superiores ao DANTE-LINEAR, para todos os valores de k. SVR obteve os melhores resultados, superior a RASSBACH-MAXENT para todos os valores de k. Todos estes resultados são estatisticamente significativos utilizando o teste de postos com sinais de Wilcoxon.

Na Figura 4.4 é apresentado NDCG@k para os métodos SVR, DANTE-SVR e RASSBACH-SVR utilizando o melhor conjunto de atributos propostos para cada método para representar os artigos. Observando as Figuras 4.3 e 4.4 percebe-se que a utilização de SVR melhorou bastante o desempenho de DANTE e ligeiramente o

desempenho do RASSBACH. Novamente, na Figura 4.4, SVR é o método que apresentou os melhores resultados sendo superior ao RASSBACH-SVR para todos os valores de k, o que pode ser atribuído à melhor escolha de atributos. Estes resultados são estatisticamente significativos.

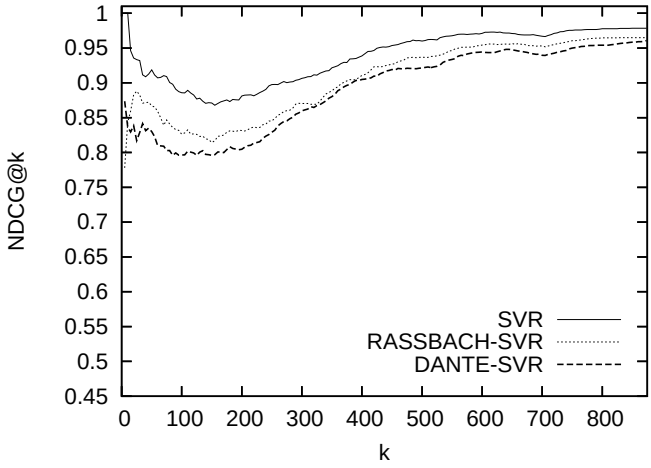

Figura 4.4. NDCG@k obtido pelos métodos DANTE-SVR, RASSBACH-SVR, e SVR.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho, foi proposta uma nova abordagem, baseada em regressão e aprendizado de máquina, para combinar indícios dos artigos de comunidades colaborativas. Esta abordagem tenta unir em um método as mais relevantes contribuições dos indícios anteriormente propostos na literatura com novos indícios apresentados aqui. Sendo assim, neste trabalho foi realizado a análise do impacto de diversos indícios ao estimar a qualidade de artigos da Wikipédia. Os indícios foram analisados dividindo-os em 6 tipos: tamanho, estrutura, estilo, legibilidade, revisão e atributos de redes complexas, sendo que os quatro primeiros são evidências do texto. A partir da análise destes grupos, foi possível observar a importância dos atributos de texto que obteveram um resultado superior aos demais tomados isoladamente. Um resultado importante, pois são os atributos de mais simples acesso em qualquer enciclopédia digital livre. Além disso, conseguiu-se resultados melhores ainda quando combinou-se indícios de texto com os de revisão e de grafo. Por outro lado, foi possível perceber que alguns atributos, como *probReview* e atributos de redes complexas, possuem um custo computacional alto e um baixo impacto ao estimar a qualidade dos artigos. Logo, estes atributos poderiam ser descartados, em princípio. Percebeu-se também a importância de observar o tamanho e a estrutura do texto ao estimar a qualidade de um artigo da Wikipédia. Além disso, constatamos uma maior dificuldade na estimativa da qualidade de artigos que possuem uma qualidade mais alta.

O método proposto neste trabalho foi o que apresentou o melhor desempenho entre todos aqueles que foram comparados.

Em trabalhos futuros, pretendemos:

- Considerar artigos com várias classes de qualidade, que foram ignorados neste trabalho;

- Utilizar uma nova e maior amostra, para efeitos de validação;
- Aplicar esta abordagem em outras bibliotecas colaborativas on-line, tais como: comunidades hospedadas na *Wikia*[1], *Planet Math*[2] e Google Knol[3];
- Analisar o impacto dos atributos utilizando alguma métrica que leve em conta a dependência dos atributos;
- Verificar se o impacto dos atributos é diferente entre artigos de alta e baixa qualidade;
- Utilização de atributos que levem em conta a qualidade das citações externas e também a qualidade das traduções dos artigos. A qualidade da citação é um indício importante para atender o critério de verificabilidade da informação na Wikipédia [Wikipedia, 2009]. Um exemplo de métrica que verifica a qualidade de citações é o fator de impacto [Garfield, 1972];
- Criar ferramentas para auxiliar no processo de classificação de qualidade da Wikipédia e identificação de vandalismos;
- Estudar a influência da qualidade dos artigos em processos de busca.

[1] http://www.wikia.org
[2] http://www.planetmath.org/
[3] http://knol.google.com

Referências Bibliográficas

Adler, T. B. & de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 261--270, Banff, Alberta, Canada.

Alexander, J. E. & Tate, M. A. (1999). *Web Wisdom; How to Evaluate and Create Information Quality on the Web*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

Barroso, M. M. A. (1998). A matemática na limpeza urbana: Trajeto Ótimo do caminhão de lixo. In *XXI CNMAC - Congresso Nacional de Matemática Aplicada e Computacional*, Caxambu, Minas Gerais, Brasil.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; Zhang, C. & Ross, K. (2008). Identifying video spammers in online social networks. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp. 45--52, Beijing, China.

Björnsson, C. (1968). *Lesbarkeit durch Lix*. Stockholm: Pedagogiskt Centrum.

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

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

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

Brown, R. (2009). Does fundamentalist religion cause the rejection of evolution? or is it the other way around? `http://karmatics.com/docs/evolution-and-wisdom-of-crowds.html`.

Cassel, R. (1995). Selection criteria for internet resources. *College and Research Libraries News*, 56(2):92--93.

Castillo, C.; Donato, D.; Gionis, A.; Murdock, V. & Silvestri, F. (2007). Know your neighbors: web spam detection using the web topology. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 423--430, Amsterdam, The Netherlands.

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

Chu, Y. (1997). Trust management for the world wide web. Master's thesis, MIT, USA.

Coleman, M. & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283--284.

Dalip, D. H.; Gonçalves, M. A.; Cristo, M. & Calado, P. (2009). Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia. In *JCDL '09: Proceedings of the 9th Joint Conference on Digital Libraries*. ACM/IEEE.

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

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

Faraco, C. E. & de Moura, F. M. (1997). *Gramática Faraco & Moura*. Editora Ática, São Paulo - SP, Brasil.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, pp. 221--235.

Fogg, B. J.; Marshall, J.; Laraki, O.; Osipovich, A.; Varma, C.; Fang, N.; Paul, J.; Rangnekar, A.; Shon, J.; Swani, P. & Treinen, M. (2001). What makes web sites credible?: a report on a large quantitative study. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 61--68, Seattle, Washington, United States.

Fogg, B. J.; Soohoo, C.; Danielson, D. R.; Marable, L.; Stanford, J. & Tauber, E. R. (2003). How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *DUX '03: Proceedings of the 2003 conference on Designing for user experiences*, pp. 1--15, San Francisco, California, USA.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(60):471--479.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):901--902.

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill International Book Co.

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

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41--48, Athens, Greece.

Korfiatis, N.; Poulos, M. & Bokos, G. (2006). Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review*, 30(3):252--262.

Krowne, A. (2003). Building a digital library the commons-based peer production way. *D-Lib magazine*, 9(1082).

McLaughlin, G. H. (1969). Smog grading: A new readability formula. *Journal of Reading*, pp. 639--646.

Mingus, B. (2008). personal communication.

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

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

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

Ressler, S. (1993). *Perspectives on electronic publishing: standards, solutions, and more*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Roberts, T. S. (2004). *Online Collaborative Learning: Theory and Practice*. Idea Group Pub, USA.

Smith, E. A. & Senter, R. J. (1967). Automated readability index. *Aerospace Medical Division*.

Stvilia, B.; Twidale, M. B.; Smith, L. C. & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality-ICIQ 2005*, pp. 442--454.

Style & Diction (2008). GNU Project. http://www.gnu.org/software/diction/.

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

Veltman, K. H. (2005). Access, claims and quality on the internet – future challenges. *Progress in informatics : PI*, 2:17–40.

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

Wikipedia (2008a). http://en.wikipedia.org/wiki/Wikipedia.

Wikipedia (2008b). Featured article candidates. http://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates.

Wikipedia (2008c). Good article nominations. http://en.wikipedia.org/wiki/Wikipedia:Good_article_nominations.

Wikipedia (2008d). Version 1.0 editorial team/assessment. http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment.

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

Wikipedia (2009). Verifiability. `http://en.wikipedia.org/wiki/Wikipedia:Verifiability`.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, pp. 80–83.

Witten, I. H. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 1ª edição.