Análise de Metadados para Inferência da Qualidade de Artigos da Wikipedia

1st Rodrigo Schmidt Nürmberg
Universidade Federal do Rio de Janeiro
Universidade Federal do Rio de Janeiro
Rio de Janeiro-RJ, Brasil
rsn86@rsn86.com
ORCID: https://orcid.org/0000-0003-0634-793X

2nd M.Sc Arlindo S. Amaral Neto
Universidade Federal do Rio de Janeiro
Universidade Federal do Rio de Janeiro
Rio de Janeiro-RJ, Brasil
arlindoneto@ufrj.br
ORCID: https://orcid.org/0000-0002-9692-3568

***Resumo*—A crescente quantidade de informações disponibilizadas na Web traz consigo uma preocupação com relação à qualidade destes dados já que cada vez mais eles são utilizados na tomada de decisões. Este trabalho analisou estatisticamente alguns metadados que podem ser utilizados para se determinar, de forma automática e objetiva, a qualidade dos artigos encontrados na Wikipedia.**

***Index Terms*—qualidade da informação, dimensões de qualidade, metadados, Wikipédia**

I. Introdução

A popularização das tecnologias da informação e da comunicação e o surgimento de ferramentas facilitadoras da publicação de informações na Web, tais como blogs, possibilitaram que o usuário deixasse de ser apenas um mero consumidor de informações e passasse a produzi-las. Estas informações muitas vezes são publicadas sem passarem por um sistema capaz de assegurar a qualidade dos dados.

A despeito disso, a Web é cada vez mais utilizada como fonte de informação. Seu uso nas tomadas de decisão gera uma preocupação crescente sobre a qualidade da informação e determiná-la tornou-se imprescindível.

Contrapondo à geração descoordenada de informação, existem iniciativas para geração colaborativa de conhecimento. Geralmente estas iniciativas utilizam ferramentas como os Wikis. Estes sistemas consistem em uma coleção de documentos Web que podem ser facilmente criados e editados através de navegadores, democratizando a publicação de informações. O mais evidente destes projetos é a Wikipedia[1], uma enciclopédia online, multilíngue, livre e colaborativa que surgiu em 2001 e atualmente conta com mais de 55 milhões de artigos[2], mantidos por voluntários de diversas regiões do mundo.

A Wikipedia conta com um sistema de revisão a fim de assegurar a qualidade de seu conteúdo e evitar a ocorrência de novos escândalos envolvendo falsas biografias e artigos tendenciosos. Porém, este é um trabalho manual e voluntariado, geralmente incapaz de lidar com a enorme quantidade de informações.

Diante deste cenário, este trabalho buscou identificar a possibilidade de se inferir a qualidade dos artigos da Wikipedia através de uma abordagem estatística, utilizando os metadados disponíveis.

[1]http://en.wikipedia.org/wiki/Wikipedia:About

[2]http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total

II. Qualidade da Informação

A falta de uma definição ontológica de qualidade ilustra a dificuldade encontrada ao se tentar tratar computacionalmente esta questão em relação de termos objetivos.

A norma ISO 9000:2005 define qualidade como: ”a totalidade de características de uma entidade que lhe confere a capacidade de satisfazer às necessidades explícitas e implícitas”. As necessidades explícitas são compostas pelas condições de utilização do produto, seus objetivos, funções e desempenho esperado. As necessidades implícitas são aquelas que, embora não expressas nos documentos do produtor, ainda assim são importantes para os usuários [3].

Esta definição relativiza a qualidade, uma vez que ela passa a depender de seus produtores e usuários. É a mesma relatividade muitas vezes encontrada na definição de qualidade da informação ([10], [11], [13]).

Para [15], a qualidade da informação é dependente da opinião e análise do seu utilizador. Sendo assim, encontrar um método computacional para mensurar e avaliar a qualidade de uma determinada informação, sem a necessidade de intervenção humana, é uma tarefa árdua.

[12] aponta que mensurar a qualidade da informação é um assunto dificultado por razões como a natureza subjetiva da necessidade do usuário, as origens da informação, a abundância de dados, entre outras. É neste contexto que os metadados aparecem como alternativa para minimizar o problema.

De acordo com a National Information Standards Organization (NISO), metadados são informações estruturadas que descrevem, explicam, localizam, ou tornam mais fácil de obter, usar ou gerir informações. [5] trata os metadados como ”dados sobre dados”. Ele complementa: ”neste contexto, metadado refere-se a alguma estrutura descritiva da informação sobre outro dado, usado para ajudar na identificação, descrição, localização e gerenciamento de recursos da Web”.

Dimensões de qualidade são compostas por uma ou mais características mensuráveis, e podem ser entendidas como aspectos pelos quais se observam a qualidade de objetos. No contexto de qualidade de informação, de acordo com [1], a

dimensão de qualidade é percebida como uma perspectiva do usuário sobre a qualidade do documento e um conjunto de dimensões de qualidade é adotado como o critério na avaliação da qualidade da informação. As dimensões de qualidade podem ser representadas, com certa imprecisão, através de metadados associados.

Neste trabalho, para mensurar as dimensões de qualidade, coletaram-se informações (metadados) relacionadas aos artigos da Wikipédia utilizados no experimento e, para melhorar a precisão desta representação, em alguns casos, foi necessário utilizar mais de um metadado para representar ou mensurar uma dimensão de qualidade.

[13] observa que "as definições sobre a qualidade da informação têm sido feitas sob o ponto de vista de definições específicas e subjetivas, seguidas por definições ad-hoc. Isso tem resultado em inúmeras interpretações pouco claras do conceito, imperfeitas e de alguma forma caóticas."Partindo do pressuposto de que a qualidade pode basear-se em atributos de valor, [14] considera a qualidade como uma das dimensões do "valor agregado" da informação; ou seja: a informação se torna mais valiosa quando é organizada, sintetizada e julgada. Ele considera qualidade as seguintes características: precisão, abrangência, atualidade, confiabilidade, validade dos dados e informações de um sistema.

III. Dimensões de Qualidade Analisadas

Com base nas definições das dimensões de qualidade descritas por [1] e [6], selecionou-se um conjunto destas dimensões relacionadas principalmente ao conteúdo do artigo, deixando de lado questões como apresentação e manutenibilidade. A seguir efetuou-se uma análise dos artigos da Wikipedia e de suas páginas auxiliares (como a página de discussões e o histórico) a fim de determinar quais metadados poderiam ser extraídos para representar as dimensões de qualidade selecionadas.

Foram utilizadas como classificação da qualidade dos artigos, o seu respectivo PageRank e quando disponível, a classificação feita pelos revisores da Wikipedia. A tabela I exibe o mapeamento identificado entre as dimensões de qualidade e metadados.

Com a crescente facilidade de se inserir conteúdo na Internet, é cada vez mais difícil inferir o nível de obsolescência da informação. É comum a utilização de regras baseadas na data de atualização e/ou publicação para medir o nível de atualidade conteúdo, do ponto de vista da provisão da informação, tais regras são questionáveis.

A atualidade refere-se ao nível de aceitação de um dado para a realização de uma determinada tarefa, ou seja, determina se os dados são suficientemente atualizados ao propósito em questão. Deste modo, é imprescindível a adoção de um contexto para determinar, precisamente, a atualidade da informação que está sendo analisada. Como os artigos analisados fazem parte de um contexto tecnológico (banco de dados) para identificar o quão atual ele era até a data do experimento foi utilizada a data da última edição do artigo para identificar o quão atual ele era até a data do experimento.

Tabela I
Mapeamento de dimensões de qualidade em metadados

Dimensão de Qualidade	***Metadados***
Acurácia	Número de erros ortográficos, número de sugestões corretas
Atualidade	Quantidade de dias desde a última edição até a data atual
Completeza	Valor do Hub de uma página, que é calculado com base nos links para os quais a página aponta
Reputação	Valor do Authority de uma página, que é calculado com base nos links que apontam para a página
Relação Informação Ruído	Número de palavras únicas, número de stop words, número de palavras

Para analisar as dimensões de completeza e reputação utilizamos a estrutura de links do ambiente web. [4] afirma que essa estrutura pode ser uma rica fonte de informações sobre o conteúdo do ambiente e também que os links entre as páginas codificam uma quantia oculta considerável de julgamento humano sobre a qualidade das páginas.

De acordo com [6], a análise da estrutura de links é a mais proeminente fonte de conhecimento usando informação encontrada fora da própria página Web e explora meramente a estrutura de links e os links entre as páginas Web para analisar a qualidade dessas páginas. A suposição básica é que o número de links que apontam para uma página é a medida para a popularidade e conseqüentemente para a qualidade dessa página e esses links são denominados in-links ou back-links. Um outro conceito utilizado para analisar a estrutura de links da web são os indegrees e outdegrees, que são respectivamente o número de páginas que apontam para a página e o número de páginas que a página aponta.

Para analisar a estrutura de links da Web este trabalho utiliza a ferramenta JUNG[3], que é uma biblioteca que fornece meios para modelar, analisar e visualizar informações que podem ser representadas através de grafos e se baseia nos conceitos de hub e authority definidos por [4] para implementar o algoritmo HITS utilizado na análise.

De acordo com [4], authorities são páginas reconhecidas como fornecedoras de informação significante, confiável e útil sobre um determinado assunto. E hubs são páginas que fornecem links úteis para páginas com conteúdo relevante, ou seja, authorities. Hubs e authorities mostram o que pode ser chamado de um relacionamento sinérgico: um bom hub é uma página que aponta pra vários bons authorities; e um bom authority é uma página que é apontada por vários bons hubs.

Para calcular os valores de hub e authority de cada página (artigo da Wikipédia) analisada, foi montado um grafo com todas as páginas da web, encontradas pelos motores de busca do Yahoo, que apontavam para aquele artigo e também com todas as páginas apontadas por este artigo, essa informação foi obtida através dos links contidos na página, com exceção dos links de navegação interna da página que foram desconsi-

[3]http://jung.sourceforge.net/

derados. Após montado o grafo, os valores de hub e authority foram calculados com o algoritmo HITS.

O conceito das duas dimensões de qualidade analisadas com base na estrutura de links da web, seguem as definições dessas dimensões de acordo com [1]. A completeza expressa se os dados disponibilizados são suficientemente completos em largura, profundidade e escopo para a tarefa em questão. A reputação expressa se os dados são confiáveis ou altamente recomendados em termos de sua origem ou conteúdo.

Percebe-se que esses conceitos se sobrepõem em alguns pontos, portanto a princípio essas duas dimensões de qualidade foram relacionadas com os conceitos de hub e authority definidos por [4], mais precisamente, completeza com o conceito de hub e reputação com o conceito de authority.

A acurácia, segundo a definição de Barros (2009), serve para indicar se os dados estão livres de erros. Nessa intenção, ela foi calculada com base na verificação ortográfica. Levou-se em conta o número de erros de grafia e também o número de sugestões apresentadas pelo verificador ortográfico.

Além destas dimensões de qualidade também foram analisadas medidas com base no arquivo, como tamanho do arquivo, comprimento da URL e do título e outras baseadas na linguagem, como número de palavras, número de stopwords e número de palavras únicas, todas estas medidas definidas em [6]. A partir destas medidas, calculou-se a relação informação ruído da seguinte forma: I/N = número de palavras únicas / número de palavras.

IV. Experimento

Uma vez determinados os metadados a serem extraídos, foram selecionados aleatoriamente 207 artigos, pertencentes à categoria Databases da Wikipedia em inglês, para compor o dataset.

Para coleta dos artigos e de seus metadados utilizou-se o seguinte conjunto de APIs:

- JWBF (Java Wiki Bot Framework): provê métodos para ler coleções de artigos, obter informações sobre a edição e a classificação dos artigos.
- Hunspell: dicionário utilizado pelos projetos Mozilla Firefox e OpenOffice para a verificação ortográfica.
- JavaPageRank: obtém o PageRank de uma URL.
- JUNG: em conjunto com buscadores Web, permite a construção de um grafo que representa a estrutura de links da Web e assim é possível obter o hub e o authority de uma dada página.

Os resultados obtidos em todas as etapas do processo foram persistidos em um banco de dados a fim de garantir que todos os metadados se referem a uma mesma versão do artigo.

Para análise dos metadados, plotou-se a média dos valores obtidos para cada um versus seu respectivo PageRank e também versus a avaliação recebida dos revisores da Wikipedia. Utilizou-se o Weka para visualização e mineração dos dados.

V. Resultados

Como um indicador da qualidade dos artigos utilizamos a revisão manual feita pelos revisores do Wikipedia e também o GoogleTM PageRank. O sistema PageRank é usado pelo motor de busca Google para ajudar a determinar a relevância ou importância de uma página. Segundo [9], o Google PageRank possui uma forte correlação com o número de visualizações de uma página da Wikipedia, por tanto, um representante útil e preciso de sua importância.

De acordo com [7], o PageRank é um método para calcular o ranking de páginas com base no mapa da Web, com aplicações em busca, navegação e estimativa de tráfego. A ordenação pelo PageRank valoriza as páginas mais importantes e centrais da Web.

Os resultados foram normalizados para facilitar a visualização e comparação entre os diversos metadados. Foi desconsiderado o fato do dataset não possuir uma distribuição homogênea de artigos em relação aos pagerank's. O gráfico 1 exibe esta distribuição. Desta forma, para alguns metadados, os picos apresentados nos gráficos foram desconsiderados.

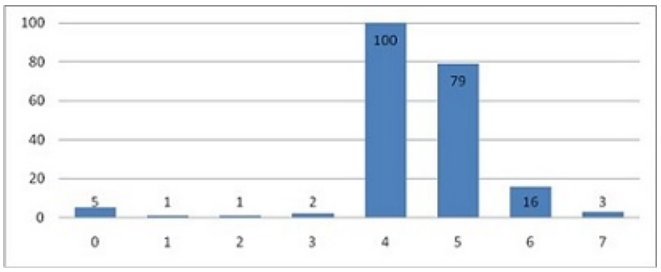

Figura 1. Quantidade de páginas por PageRank.

Ao plotar a avaliação recebida dos revisores da Wikipedia contra o PageRank, observamos que ambos possuem crescimento ascendente a partir do PageRank 4. Não foi possível verificar o crescimento em todo o gráfico devido à baixa quantidade de páginas analisadas com PageRank entre 1 e 3, o que facilitou a análise dos dados, pois tornou desnecessária a plotagem dos metadados contra ambas.

Os valores de assessments foram gerados a partir da avaliação dos revisores da Wikipedia e convencionados em notas que variam de 0 a 5 no universo das páginas avaliadas. Deste modo, os demais gráficos foram plotados apenas contra o PageRank do Google. O gráfico 2 apresenta esta curva.

O gráfico 3 ilustra o relacionamento sinérgico entre Hub e Authority. Observa-se que as páginas com PageRank igual a 4 ou igual a 5 são as que melhor representam esse tipo de relacionamento, pois ilustra o comportamento desejável para esse relacionamento. Hub e authority, para seu respectivo PageRank, devem ser proporcionais e o gráfico ideal deve manter proporcionalidade crescente.

Os resultados encontrados as curvas de Outdegree e Indegree, apresentadas no gráfico 4, servem como complemento às curvas de Hub e Authority, respectivamente. Percebe-se que as curvas possuem uma tendência de crescimento em relação ao Pagerank. Isso indica que o resultado do gráfico

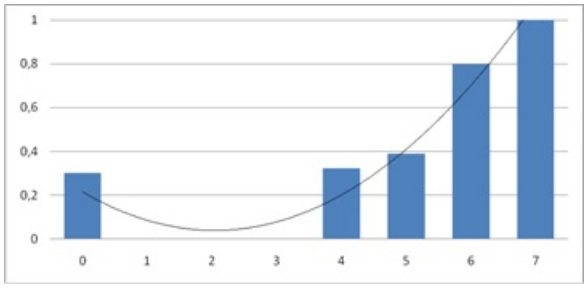

Figura 2. Crescimento do Assessment em relação ao PageRank e linha de tendência.

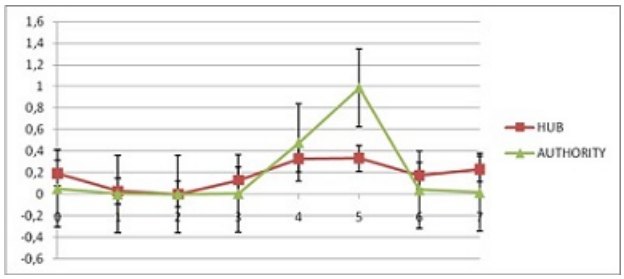

Figura 3. Relacionamento Sinérgico Hub/Authority.

3 foi penalizado por causa do dataset utilizado, uma vez que o comportamento ideal é uma curva exponencial crescente, tanto para a curva de outdegree quanto para a de indegree.

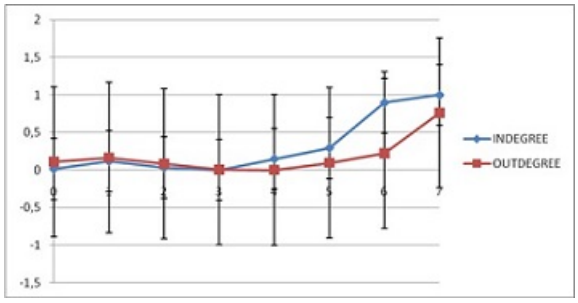

Figura 4. Curvas de Indegree e Outdegree.

A curva de atualidade, apresentada no gráfico 5, também sofreu influência do dataset utilizado. A curva ideal para ilustrar esse metadado deveria decrescer, tendendo a zero, ao passo que o Pagerank aumenta. Neste estudo, este foi o metadado que apresentou o maior desvio padrão entre os demais analisados.

Para analisar a relação sinal ruído de uma página foram levados em consideração três metadados: quantidade palavras, de Stop Words, de palavras únicas. Para verificarmos o comportamento ideal dessa relação, a curva de Stop Words deve decrescer exponencialmente com o aumento do PageRank, ao contrário das curvas de quantidade palavras e palavras únicas que devem crescer exponencialmente. O gráfico 6 apresenta as curvas obtidas para a relação sinal ruído.

A acurácia de um artigo foi aferida com base no número de erros e de sugestões aceitas. O gráfico 7 mostra as curvas obtidas para a relação de acurácia do conteúdo das páginas analisadas. Tal gráfico deveria apresentar comportamento contrário ao encontrado: as curvas deveriam tender a zero com o aumento do PageRank.

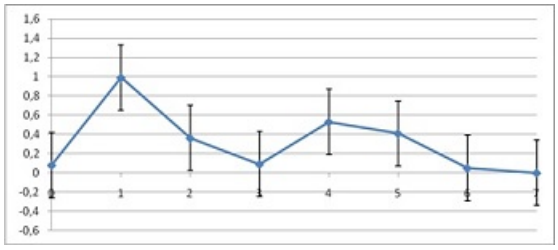

Figura 5. Curva de Atualidade.

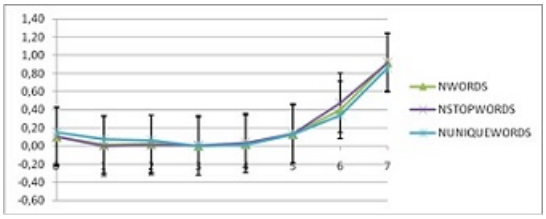

Figura 6. Relação Sinal Ruído.

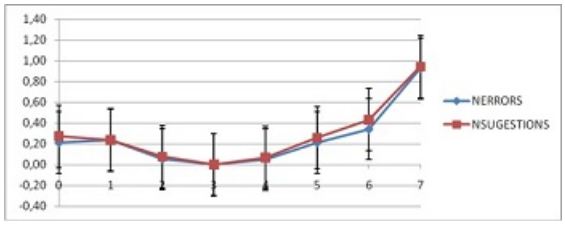

Figura 7. Acurácia do conteúdo das página.

A tabela II sumariza os resultados obtidos durante a realização do experimento.

Tabela II
RESULTADOS DA ANÁLISE

Google Page Rank	Quantidade de Páginas	Atualidade	Hub	Authority	Indegree	Outdegree	Words	Stop Words	Unique Words	Errors	Sugestions	Assessment
0	5	0,08	0,19	0,05	0,01	0,11	0,11	0,10	0,15	0,21	0,27	0,30
1	1	0,99	0,03	0,00	0,12	0,17	0,02	0,00	0,07	0,24	0,24	0,00
2	1	0,36	0,00	0,00	0,03	0,09	0,02	0,01	0,06	0,05	0,08	0,00
3	2	0,09	0,13	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00
4	100	0,53	0,33	0,48	0,15	0,00	0,02	0,03	0,01	0,05	0,07	0,32
5	79	0,41	0,33	0,99	0,29	0,09	0,13	0,13	0,13	0,21	0,26	0,39
6	16	0,05	0,17	0,04	0,90	0,22	0,40	0,47	0,34	0,34	0,43	0,80
7	3	0,00	0,23	0,02	1,00	0,76	0,92	0,92	0,86	0,93	0,94	1,00
Desvio Padrão	39,97	0,34	0,12	0,36	0,41	0,25	0,32	0,33	0,29	0,30	0,30	0,38

VI. Conclusões e Trabalhos Futuros

A abordagem estatística utilizada neste trabalho mostrou que, apesar dos poucos artigos avaliados, é possível se inferir a qualidade de artigos da Wikipedia através de metadados. Em geral os metadados analisados refletem bem a qualidade dos artigos em questão. Ressaltando-se que o PageRank foi adotado como indicador da qualidade por ter apresentado comportamento semelhante a avaliação manual feita pelos revisores da Wikipedia. Uma alternativa seria trabalhar com uma avaliação manual e confiável, feita por especialistas.

Uma vez que apenas a data da última edição foi considera, perdeu-se a natureza temporal dos Wikis, outros aspectos também ignorados foram o número de edições e os usuários que fizeram as edições. Talvez este seja um motivo para que a atualidade tenha apresentado o pior resultado dos metadados avaliados.

Pode-se perceber com bases em dados preliminares e trabalhos relacionados ([2], [9]) que a qualidade emerge da colaboração e apesar do número de edições estar relacionado a qualidade de um artigo, a qualidade das edições e o perfil dos usuários que as fizeram também devem ser levados em consideração. Uma exploração adequada destes aspectos pode levar a modelos melhores para inferência da qualidade em tais ambientes.

Uma abordagem que se mostrou promissora e precisa ser testada no ambiente colaborativo dos Wikis é o modelo fuzzy elaborado por [1], uma vez que ele é capaz de lidar melhor com as imprecisões da qualidade.

Referências

[1] BARROS, Ricardo O. (2009) QUALIDADE DE INFORMAÇÃO NA WEB: UM PROGNÓSTICO FUZZY BASEADO EM METADADOS. 2009. Tese (Doutorado em Engenharia de Sistemas e Computação) - Universidade Federal do Rio de Janeiro. Co-Orientador: Geraldo Bonorino Xexéo.

[2] DRUCK, Gregory; MIKLAU, Gerome, MCCALLUM, Andrew. (2008) Learning to Predict the Quality of Contributions to Wikipedia. In AAAI Workshop on Wikipedia and AI, 2008.

[3] ISO 9000:2005, 2005, Quality management systems - Fundamentals and vocabulary.

[4] KLEINBERG, J. (1997) Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[5] LINO, Manuel Rosa de Oliveira. Um modelo para medir a qualidade da informação de sites utilizando programação difusa / Manuel Rosa de Oliveira Lino; orientador João Bosco da Mota Alves; co-orientador Rogério Cid Bastos - Florianópolis, 2006. 115 f. Tese (Doutorado) – Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção, 2006.

[6] MANDL, Thomas (2008): Automatic Quality Assessment for Internet Pages. In: Cale-ro, Coral; Moraga, Ángeles; Piattini, Mario (eds.): Handbook of Research on Web Infor¬mation Systems Quality. Idea Group Reference: Hershey et al. S. 104-112.

[7] PAGE, Lawrence; BRIN, Sergey; MOTWANI, Rajeev; WINOGRAD, Terry. The PageRank Citation Ranking: Bringing Order to the Web. 1998.

[8] WIKIPÉDIA. Desenvolvido pela Wikimedia Foundation. Apresenta conteúdo enciclopédico. Disponível em: ¡ https://pt.wikipedia.org/wiki/Wikipédia¿. Acesso em: Out 2020.

[9] WILKINSON, Dennis M.; HUBERMAN, Bernardo A. Cooperation and quality in wikipedia, Proceedings of the 2007 international symposium on Wikis, p.157-164, October 21-25, 2007, Montreal, Quebec, Canada.

[10] WAGNER, G. The value and the quality of information: the need for a theoretical syntesis. In: WORMELL, I. (Ed.). Information quality: definitions and dimentions, London: Taylor Graham, 1990. P. 69-72.

[11] SCHWUCHOW, W. Problems in evaluating the quality of information services. In: WORMELL, I. (Ed.). Information quality: definitions and dimensions. London: Taylor Graham, 1990. P.69-72.

[12] NAUMANN, F.; ROLKER, C.; Assessment methods for information quality criteria. German research society, Berlin, 2000.

[13] WORMELL, I. (Ed.). lnformation quality: definitions and dimentions, London: Taylor Graham, 1990. Introduction, p.1-6.

[14] TAYLOR, R. S. Information values in decision contexts. Information Management Review, v. 1 n.1,p.7-55, Summer 1985.

[15] DEMING, William E. Qualidade: a revolução da administração. Tradução de Clave Comunicações e Recursos Humanos. 1. ed. Rio de Janeiro: Marques-Saraiva, 1990.