FERP: Ferramenta para Estimativa de Revocação e Precisão

Juliana Bonato dos Santos, Raquel Kolitski Stasiu, Carlos A. Heuser

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{jbonato|rkstasiu|heuser}@inf.ufrgs.br

Abstract. *Vague queries mechanisms use similarity metrics to determine the similarity between two objects which inexact matching. One of the problems of this approach is how to define an adequate threshold to be used by these queries, since that this value varies in accordance with the similarity metric used. This work presents a Web tool to estimate the quality of vague queries results for different thresholds, through recall and precision measures. This estimative allow to define an adequate threshold and similarity metrics for determined data domain.*

Resumo. *Mecanismos de consultas vagas utilizam métricas de similaridade para determinar a semelhança entre dois objetos que não possuem identificação exata. Um dos problemas dessa abordagem é a definição de um valor de corte adequado a ser utilizado por essas consultas, visto que esse valor varia de acordo com a métrica de similaridade utilizada. Esse trabalho apresenta uma ferramenta Web que estima a qualidade dos resultados de consultas vagas para diferentes valores de corte, através de medidas de revocação e precisão. Essa estimativa permite que sejam definidos os valores de corte e métricas de similaridade mais adequados para determinado domínio de dados.*

1. Introdução

Alguns bancos de dados relacionais apresentam dados com variações tanto ortográficas como de formato. Essas variações ocorrem, principalmente, em dados digitados por usuários ou provenientes da integração de sistemas [Stasiu 2005]. Por exemplo, suponhamos uma base de cidades que contenha os seguintes nomes de cidades: “Curitiba” e “CTBA”. Embora esses nomes estejam escritos de formas diferentes, ambos fazem referência à mesma cidade. Ao submeter uma consulta sobre essa base de cidades, tendo como predicado o nome dessa cidade, o processador de consulta deveria recuperar todas as representações desse objeto, independente das variações encontradas. Entretanto, para os mecanismos tradicionais de consulta, os quais utilizam a busca por coincidência exata, os nomes “Curitiba” e “CTBA” seriam considerados objetos diferentes.

Uma solução para esse problema seria a utilização de métricas de similaridade [Cohen 2000], amplamente estudadas pela área de Recuperação de Informações (RI) [Baeza-Yates e Ribeito-Neto 1999]. Essa abordagem retorna uma lista (*ranking*) com todos os elementos armazenados ordenados de acordo com o seu valor de similaridade em relação ao objeto consulta. Cabe ao usuário escolher quais dos objetos recuperados ele considera mais adequados ou fazer um refinamento na busca.

Porém, em bancos de dados relacionais, o retorno de todos os objetos armazenados sob a forma de um *ranking* é impraticável, visto que o processamento da consulta ocorre sem a intervenção do usuário. Além disso, o retorno de todos os elementos prejudica o desempenho do sistema. Portanto, é preciso definir um valor de corte (*threshold*) para reduzir o volume de dados retornados pelas consultas vagas e evitar a necessidade de intervenção do usuário durante a execução da consulta.

Para auxiliar na definição de um valor de corte adequado pode ser realizada uma avaliação da qualidade do resultado de consultas vagas. Essa avaliação permite medir quanto do resultado esperado foi alcançado para determinado valor de corte, auxiliando, assim, na definição de um valor de corte mais adequado dependendo da métrica de similaridade utilizada e do domínio dos dados consultados. Para avaliar o resultado de uma pesquisa por similaridade, duas medidas usadas na área de RI e que são utilizadas nesse trabalho são: revocação (*recall*) e precisão (*precision*). Conforme definido em [Stasiu, Heuser e Silva 2004] é possível estimar valores de revocação e precisão com pouca intervenção do usuário, através da utilização de amostras de bases de dados.

Esse trabalho apresenta uma ferramenta Web que implementa o processo de estimativa de revocação e precisão descrito em [Stasiu, Heuser e Silva 2004]. A partir dessa ferramenta pode-se avaliar a qualidade dos resultados produzidos por consultas vagas para diferentes valores de corte e métricas de similaridade. Além disso, torna-se possível a extração de parâmetros que podem ser utilizados como metadados na otimização do plano de execução de consultas vagas sobre banco de dados relacionais.

Esse trabalho está organizado da seguinte forma. Na Seção 2 encontra-se descrito o processo de estimativa de revocação. A Seção 3 apresenta os módulos da ferramenta. Na Seção 4 são apresentadas as técnicas de validação do processo de estimativa utilizadas. Na Seção 5 são apresentados os detalhes técnicos da ferramenta. Na Seção 6 são apresentadas as considerações finais.

2. Processo de Estimativa

O processo de estimativa envolve as etapas descritas a seguir [Stasiu 2005]:

1) Geração de amostras a partir dos elementos contidos em uma coluna de um banco de dados relacional (BD).

2) Para cada amostra, o usuário informa o número de objetos distintos.

3) Métodos de agrupamento, utilizando diferentes métricas de similaridade, agrupam as instâncias semelhantes de cada amostra. Esse processo é repetido até que o número de grupos gerados seja igual ao número de objetos distintos informadas pelo usuário. Cada grupo formado deve representar as variações de um mesmo objeto real.

4) Cada instância de uma amostra é utilizada como objeto consulta sobre um conjunto de dados que contém todos os elementos da amostra na qual essa instância faz parte. As consultas são realizadas para diferentes valores de corte. Os elementos contidos no grupo em que o objeto consulta está localizado são considerados os elementos que devem ser recuperados pela consulta. Essa informação é utilizada para estimar os valores de revocação e precisão.

5) São estimados valores de revocação e precisão[1] para cada valor de corte utilizado nas consultas. A partir das informações extraídas do processo de agrupamento podem ser estimados valores de revocação e precisão sem a intervenção do usuário. Metadados como a métrica de similaridade utilizada, valores de revocação e precisão e valores de corte podem ser armazenados junto com as informações estatísticas de um banco de dados relacional.

3. Módulos da Ferramenta

A seguir estão descritos os principais módulos da ferramenta. Maiores detalhes podem ser encontrados em [Bonato 2005].

3.1. Geração de Amostras

O módulo de geração de amostras tem como objetivo extrair amostras de uma base de dados a partir de um arquivo XML (*Extensible Markup Language*) contendo a extração de uma coluna de um banco de dados relacional. Essas amostras são utilizadas pelos demais módulos da ferramenta para estimar os valores de revocação e precisão.

Na ferramenta, foram implementadas duas das principais técnicas de amostragem: **aleatória** e **catação**. Durante a geração das amostras pode-se, ainda, considerar ou não **sobreposição** e **repetição** de elementos. Permitir sobreposição de elementos significa que o mesmo elemento pode aparecer mais de uma vez na mesma amostra. Permitir repetição de elementos significa que uma mesma amostra pode possuir elementos exatamente iguais.

3.2. Agrupamento

O módulo de agrupamento é responsável pelo agrupamento dos objetos semelhantes de cada amostra em grupos. O módulo de agrupamento implementa os seguintes métodos de agrupamento: **Ligação Simples**, **Ligação Completa**, **Ligação Mediana** e **Ward**. Maiores detalhes sobre esses métodos podem ser encontrados em [Wives 2004].

Os métodos de agrupamento utilizam métricas de similaridade para medir o grau de semelhança entre os objetos [Bonato 2005]. As métricas implementadas nesta ferramenta são: *Edit Distance*, *N-Grams*, *Acronyms*, *Guth*, *dateSimilarity*, *Jaro*, *Jaro-Winkler* e *Jaccard*.

Para cada amostra e métrica de similaridade, o processo de agrupamento é repetido até que o número de grupos gerados seja igual ao número de objetos distintos contidos em cada amostra, segundo parâmetros fornecidos pelo usuário. Um exemplo dos grupos gerados pela ferramenta pode ser visto na figura 1.

---

[1] Conforme [Baeza-Yates e Ribeito-Neto 1999], as medidas de revocação e precisão são definidas como:

Revocação: $\frac{|RA|}{R}$ Precisão: $\frac{|RA|}{A}$, sendo que **RA** correspondente ao número de elementos relevantes retornados pela pesquisa, **R** correspondente ao número elementos no grupo em que está localizado o objeto de consulta e **A** correspondente ao número de elementos retornados pela pesquisa. Portanto, revocação uma medida que informa o percentual de elementos relevantes que foram recuperados, e precisão é uma medida que informa o percentual de elementos recuperados que são relevantes.

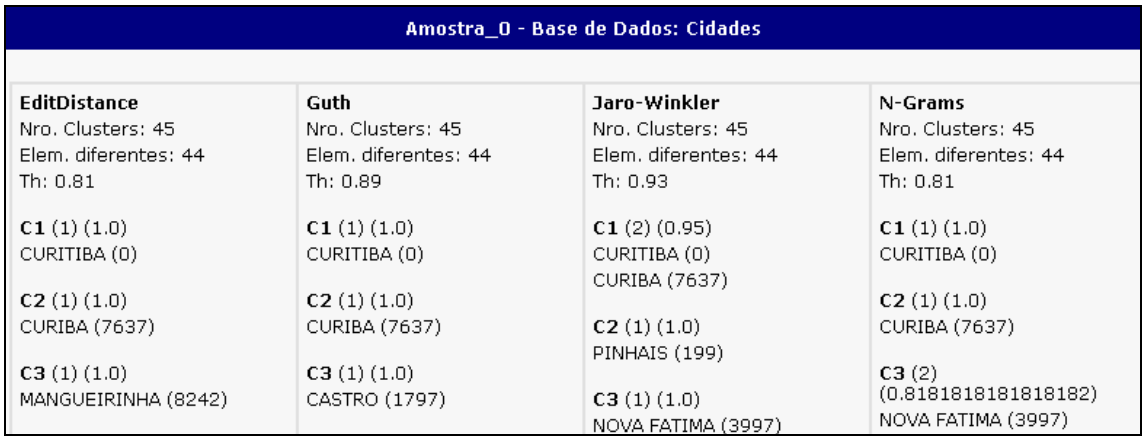

Figura 1: Grupos gerados para diferentes métricas de similaridade

3.3. Cálculo de Revocação e Precisão

Esse módulo é responsável pela estimativa de valores de revocação e precisão. Para cada amostra e métrica de similaridade, a ferramenta processa consultas utilizando cada elemento da amostra como objeto consulta. As consultas são processadas sobre a coleção de dados formada pelos elementos pertencentes à amostra cujo objeto consulta faz parte.

Após processar as consultas, a ferramenta calcula os valores de revocação e precisão. Para calcular esses valores sem a necessidade de intervenção do usuário são utilizados os grupos gerados no módulo de agrupamento. Em seguida, são calculados os valores médios de revocação e precisão para cada valor de corte, para uma determinada amostra e utilizando uma determinada métrica de similaridade.

3.4. Apresentação dos Resultados

Esse módulo é responsável pela apresentação dos resultados, gerados pelos demais módulos, ao usuário. Após calcular os valores médios de revocação e precisão são geradas representações gráficas conhecidas como curvas de revocação e precisão que permitem analisar a qualidade dos resultados de consultas vagas (figura 2).

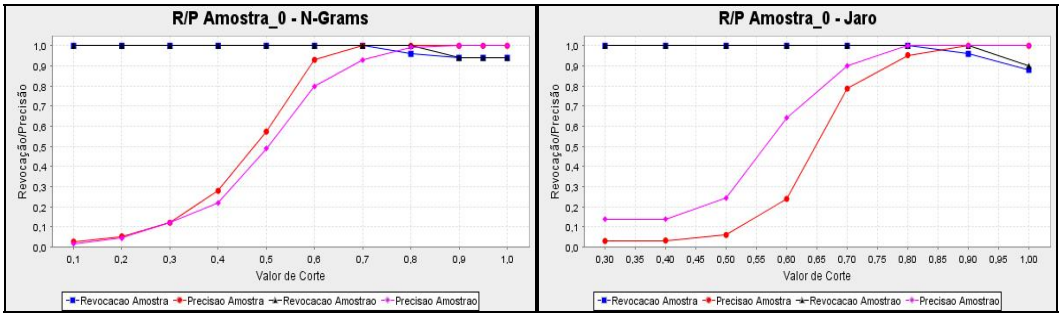

Figura 2: Gráficos de curvas de revocação e precisão

Observa-se que nos gráficos da figura 2 o eixo **x** corresponde aos valores de corte estabelecidos e o eixo **y** corresponde aos valores médios de revocação e precisão para cada valor de corte.

4. Validação do Processo de Estimativa

A validação do processo de estimativa de revocação e precisão a partir dos valores obtidos para base de dados não foi possível devido a questões de desempenho, visto que se tornava inviável a realização de todo o processo para bases de dados com mais de 500 elementos. Assim, optou-se pela geração de uma amostra contendo um número significativo de elementos para que o processo de estimativa possa ser validado. O termo *amostrão* está sendo utilizado nesse trabalho para designar essa amostra grande como alternativa ao uso de todos os elementos da base.

O processo é realizado sobre essa amostra utilizando o mesmo método de agrupamento e as mesmas métricas de similaridades utilizadas para as demais amostras. Os valores de revocação e precisão obtidos são utilizados para validar o processo de estimativa. Esses valores são apresentados nos gráficos de curvas de revocação e precisão, como pode ser visto na figura 2. Observa-se nessa figura, que as curvas de revocação e precisão para a amostra e para o *amostrão* são muito parecidas.

Para medir a proximidade dos valores obtidos com o *amostrão* e os valores estimados para as demais amostras foi calculado o Desvio Quadrático Médio definido pela fórmula [Stasiu 2005]: $f\left(x_M^b, x_M^s\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{M,i}^b - x_{M,i}^s\right)^2$, em que, $x_M^b = \left(x_{M,1}^b, x_{M,2}^b, ..., x_{M,n}^b\right)$ é o valor de similaridade do *amostrão* e $x_M^s = \left(x_{M,1}^s, x_{M,2}^s, ..., x_{M,n}^s\right)$ é o valor de similaridade da amostra.

Essa medida representa a distância média entre as curvas obtidas na estimativa de revocação e precisão sobre as amostras e as curvas obtidas sobre o amostrão. Quanto menor a distância entre as curvas, mais próximas elas estão. Exemplos de gráficos de desvio quadrático podem ser visto na figura 3.

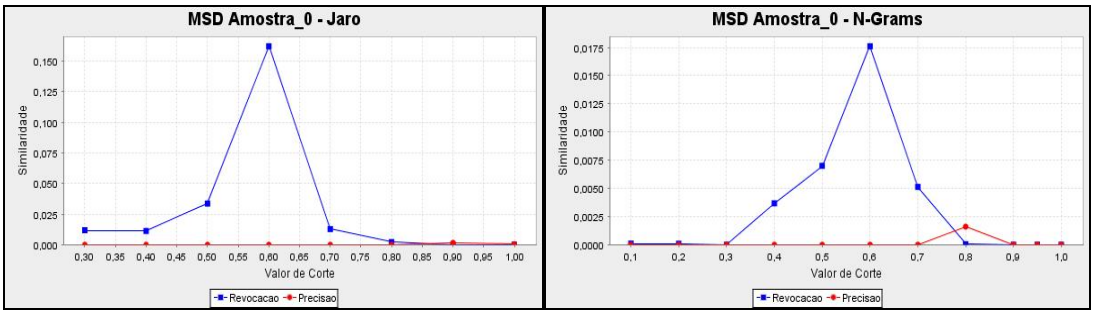

Figura 3: Gráficos de desvio quadrático

5. Aspectos técnicos da Ferramenta

A ferramenta foi implementada utilizando as linguagens de programação Java e JSP (*JavaServer Pages*), podendo, portanto, ser acessada através da Web. Como servidor de aplicações Java para Web é utilizado o Tomcat. Para a geração dos gráficos é utilizada a biblioteca JfreeChart. Os resultados são armazenados no formato CSV (*Comma Separated Value*) e XML. Por se tratar de uma ferramenta que tende a evoluir, a arquitetura prevê a inclusão de novas métricas de similaridade e métodos de agrupamento, permitindo o reaproveitamento dos demais módulos. Mais detalhes sobre o projeto e implementação do sistema está disponível em [Bonato 2005].

6. Considerações Finais

A ferramenta apresentada nesse trabalho implementa um processo de estimativa de valores de revocação e precisão através de amostras extraídas de um banco de dados, com pouca intervenção do usuário. Embora a ferramenta não vise uma integração direta com um SGBD, os valores gerados pela ferramenta podem ser utilizados como metadados e assemelham-se aos dados estatísticos gerados por bancos de dados relacionais. Cabe ao usuário analisar os resultados e, assim, extrair previamente os metadados mais adequados para que o processador possa utilizar na execução de consultas.

A partir de uma análise desses valores pode-se definir um valor de corte adequado para determinado domínio de dados. A definição de um valor de corte adequado reduz o tempo e o custo na execução de consultas vagas. Além disso, elimina a propagação de resultados incorretos entre os operadores utilizados no processamento dessas consultas.

A geração de metadados, bem como a utilização dos mesmos no processamento de consultas vagas, são temas que ainda estão sendo pesquisados. Portanto, a ferramenta desenvolvida nesse trabalho permite, também, uma maior praticidade aos usuários na realização de experimentos. Outra funcionalidade interessante da ferramenta é a possibilidade de análise e validação dos valores estimados através de métodos estatísticos básicos, como o desvio médio quadrático.

Referências Bibliográficas

Baeza-Yates, R. e Ribeiro-Neto, B. (1999) “Modern Information Retrieval”, New York, ACM Press.

Bonato, J. S. (2005) “Ferramenta para Estimativa de Revocação e Precisão usando Amostras de Banco de Dados”, Projeto de Diplomação, Porto Alegre, Instituto de Informática, UFRGS.

Coehn, W. W. (2000) “Data integration using similarity joins and a word-based information representation languague”, ACM Trans. Inf. Syst, v.18, n. 3, pág.288-321.

Stasiu, R. K.; Heuser, C. A.; Silva, R. (2004) “Estimating recall and precision for imprecise queries in databases”, Relatório de Pesquisa 348. Porto Alegre, Instituto de Informática/UFRGS.

Stasiu, R. K. (2005) “Processamento de consultas vagas em banco de dados” 56 f. Proposta de Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

Wives, L. K. (2004) “Utilizando Conceitos como descritores de Textos para o processo de identificação de conglomerados (clustering) de documentos.” 136 f. Tese (Doutorado em Ciência da Computação) - Instituto de Informática, UFRGS, Porto Alegre.