Recuperação de Informações para Bases de Documentos Estruturados

Um Modelo de Inferência Automática das Classes Estruturais e de suas Relevâncias

Ruiter Braga Caldas[1], **Horácio Fernandes de Oliveira**[1]

Universidade Federal de Minas Gerais[1]
Instituto de Ciências exatas
Departamento de Ciência da Computação

{Ruiter,Horácio}@Dcc.Ufmg.Br

Resumo. *A presença de diferentes itens de informação em um único documento, tais como títulos, resenhas, parágrafos, etc, é uma característica dos documentos ditos estruturados ou semi-estruturados. Nas páginas da Internet, é comum encontrarmos ainda espaços publicitários, notas de identificação de empresas, menus e diversos outros itens de informação. Pesquisa recente demonstra uma metodologia de exploração da importância da estrutura dos documentos de uma base a fim de melhorar a qualidade das respostas dos sistemas de recuperação de informação. Porém a identificação das classes estruturais nos documentos ainda é realizada de forma manual. Neste trabalho, é proposto um modelo de inferência automática das classes estruturais, bem como a atribuição de um grau de importância para cada uma utilizando a formatação existente nos mesmos, resultando, assim, em uma praticidade muito maior para a recuperação de informações em documentos estruturados na Internet.*

1 Introdução

A Internet trouxe inúmeras oportunidades para a comunicação em massa, tanto na capacidade de difusão como em volume de informação disponibilizada, com isso aumentou o interesse na área de processamento, armazenamento e recuperação de informações. Em princípio os sistemas computacionais foram criados para serem usados exclusivamente como mecanismos de armazenamento e recuperação de dados locais, hoje com o crescimento constante dessa grande rede de comunicação, com uma enorme quantidade de informações distribuídas e a inerente desordem destas informações, inúmeros modelos tecnológicos de recuperação de informações vem sendo desenvolvidos com o objetivo de facilitar a tarefa de caminhar por esta teia.

De forma genérica e sucinta, dada uma consulta formada por palavras chaves, os sistemas de recuperação de informações fazem uma análise da similaridade de cada documento de uma base com a consulta formulada, de forma a selecionar os documentos que mais provavelmente atendem a consulta. No entanto, de uma forma geral, os sistemas de inferência de similaridade consideram cada documento como um elemento indivisível, ainda que em muitas ocasiões cada documento seja composto por inúmeros itens de informações que, se identificados e bem trabalhados, podem trazer melhorias consideráveis à qualidade de respostas dos sistemas referenciados [6].

Em seu trabalho, Oliveira [6] propõe uma metodologia de exploração da importância da estrutura dos documentos de uma base a fim de prover uma melhor qualidade de respostas aos sistemas de recuperação de informações. No trabalho citado, a identificação destas classes estruturais nos documentos web é realizada de forma manual. No presente trabalho, é proposto um modelo de inferência automática dessas classes estruturais, bem como a importância de cada uma, em documentos web através da formatação existente nestes últimos, resultando, assim, em uma praticidade muito maior para a recuperação de informações em documentos estruturados na Internet.

1.1 Trabalhos Relacionados

O principal trabalho relacionado é a dissertação de mestrado defendida por Oliveira [6], que serviu como base deste artigo e onde o aproveitamento da estrutura implícita dos documentos web é proposto. Este trabalho será melhor explicado na Seção 3.

Kotsakis [4] propôs uma metodologia de recuperação de informação em bases de documentos XML. É sugerida a criação de um peso a cada *path* da árvore XML. No entanto, tais pesos são estipulados pelo responsável de cada documento, sem uma proposta de inferência automática.

No trabalho apresentado por Al-Khalifa *et al* [1], é proposta uma metodologia de inclusão de ranking de relevância aos processos de buscas em documentos XML. Para tanto, é definida uma álgebra relacional, que viabiliza o proposto. Também não há inferência automática.

1.2 Organização do Artigo

Este artigo está organizado como segue: serão apresentados alguns conceitos básicos na Seção 2. O trabalho em que este artigo se baseia será explicado na Seção 3. O modelo proposto neste artigo será apresentado na Seção 4. A explicação da implementação realizada e os resultados obtidos são apresentados na Seção 5. A conclusão e trabalhos futuros na Seção 6.

2 Conceitos Básicos

Para a realização deste trabalho, diversos conceitos e tópicos relacionados à recuperação de informação foram estudados. A seguir será feita uma breve introdução aos conceitos considerados mais importantes.

2.1 Sistemas de Recuperação de Informação-SRI

Os SRIs surgiram em meados dos anos 40 para atender a necessidade de se extrair informação em bases de dados não estruturadas, geralmente grandes coleções de documentos textuais e bibliográficos. Devido ao crescimento explosivo da *World Wide Web*, os SRIs vem atraindo o interesse da comunidade científica. Isso se deve à necessidade cada vez maior de recuperação de informação em grandes bases de documentos de diversas áreas do conhecimento humano, tais como, medicina, história, geografia, arqueologia e biologia. As pesquisas na área de recuperação de informação são abrangentes e abordam temas como estruturas de dados e algoritmos, tratamento de consultas precisas e imprecisas, modelos, técnicas de processamento de cadeias de caracteres, interfaces visuais, entre outros[5]. Uma visão global de trabalhos nessa área pode ser encontrada em [2].

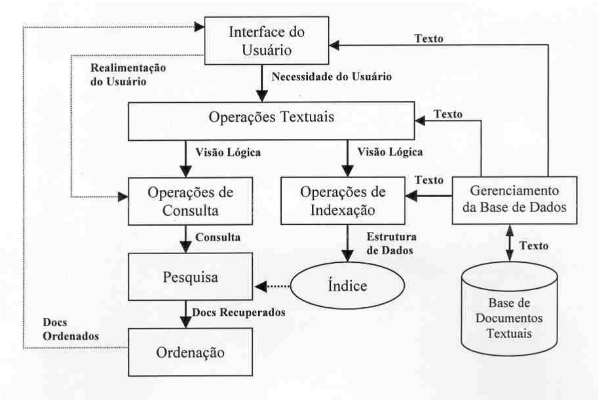

FIGURA 1: *Esquema global de funcionamento de um Sistema de Recuperação de Informação.*

A Figura 1 mostra o esquema global de funcionamento de um SRI, no qual se destacam os seguintes processos básicos [5]:

- Processo de Indexação: envolve a criação de estruturas de dados associadas aos documentos de uma coleção.
- Processo de Definição da Consulta: envolve a especificação de um conjunto de termos, associados ou não por operadores booleanos, que representa as necessidades de informação do usuário.
- Processo de Recuperação: corresponde à criação de uma lista de documentos relevantes a recuperar como resposta à consulta do usuário, baseada no cálculo do grau de similaridade entre o conteúdo dos documentos da coleção (armazenados em um índice) e os termos da consulta.

Vários modelos para RI em bases de dados textuais têm sido propostos na literatura. Será apresentado, a seguir, uma visão geral de um dos modelos mais divulgados atualmente na web [5].

2.2 Modelo Vetorial

O modelo vetorial, proposto originalmente por Salton [8], tem sido utilizado em diversos projetos de SRI pela sua simplicidade conceitual e eficácia em recuperar documentos em coleções genéricas. Neste modelo tanto os documentos quanto as consultas são representados como vetores n dimensionais dos termos da coleção. O valor das coordenadas representam a importância, ou peso, de cada termo da coleção para o documento ou consulta. A informação recuperada de um documento D de uma coleção V em resposta a uma consulta Q submetida a V, é definida pela ordem de relevância de D, calculada através do conceito de similaridade entre o vetor de termos da consulta Q e o vetor de termos do documento D, conforme descrito a seguir:

- Seja V um vocabulário formado pelo conjunto de n termos. Então, $V = \{t_1, t_2, ..., t_n\}$ representará o vocabulário V num espaço vetorial de n dimensões, onde cada termo representa uma dimensão ortogonal deste espaço.
- Seja D um documento de V. Então, o vetor $D = \{(t_1, dp_1), (t_2, dp_2), ..., (t_n, dp_n)\}$ representará o documento D neste espaço vetorial, sendo dp_j o peso (ou valor da coordenada t_j) do termo t_j no documento D.
- Seja Q uma consulta submetida a V. Então, o vetor $Q = \{(t_1, qp_1), ..., (t_n, qp_n)\}$ representará a consulta Q neste espaço vetorial, sendo qp_j o peso (ou valor da coordenada t_j) do termo t_j na consulta Q.

A similaridade entre a consulta Q e o documento D da coleção de documentos V pode ser calculada através do coseno do ângulo formado pelos vetores de termos de Q e D, usando a seguinte expressão:

$$\begin{aligned} Sim(Q, D) &= Coseno(D, Q) = \frac{D \times Q}{|D| \times |Q|} \\ &= \frac{\sum_{i=1}^{n} dp_i \times qp_i}{\sqrt{\sum_{i=1}^{2} (dp_i)^2} \sqrt{\sum_{i=1}^{2} (qp_i)^2}} \end{aligned}$$

Para o cálculo dos pesos qp_i e dp_i, utiliza-se uma técnica que faz o balanceamento entre as características do documento, utilizando o conceito de freqüência de um termo. Seja S uma consulta Q ou um documento D e $t \in S$ um termo genérico. Se a coleção possui N documentos, onde N_t e a quantidade de documentos que possui o termo t, então o inverso da freqüência do termo na coleção, ou *idf* (*inverse document frequency*) é dado por $idf(t, S) = \log \frac{N}{N_t}$. Seja $tf(t, S)$ a freqüência do termo $t \in S$, pode-se calcular o peso utilizando a seguinte fórmula: $p_i = tf(t, S) * idf(t, S)$, ou seja, é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção. As principais vantagens do modelo vetorial são sua simplicidade e a facilidade que ele provê para computar similaridades com eficiência, além do fato do modelo se comportar bem com coleções genéricas e a introdução do conceito de *matching* parcial.

2.3 Precisão X Revocação

As curvas de precisão e revocação representam a metodologia mais comumente empregada para se avaliar um SRI e foram propostas por Kent *et al* [3]. Tais conceitos representam métricas de avaliação que permitem medir a habilidade de um dado sistema de evitar os documentos não relevantes (métrica de precisão) e, ao mesmo tempo, de recuperar os documentos relevantes (métrica de revocação).

3 Modelo de RI para Documentos Estruturados

Em seu trabalho, Oliveira [6], propõe uma metodologia que permite aos SRI fazer inferência em torno da estrutura de cada documento de uma determinada base, a fim de prover uma melhor qualidade de respostas em relação àqueles sistemas que ignoram a disposição estrutural dos documentos.

A fim de melhor elucidar a abordagem utilizada, considere cada unidade de informação dos documentos como sendo uma classe estrutural. Por exemplo, considerando uma página de notícias, título é uma classe estrutural, resumo, conteúdo, menu e espaços publicitários formam outras classes estruturais. A cada uma destas classes um número pode ser atribuído como sendo o nível de relevância desta em relação ao documento. As classes que tiverem maior número de relevância são as mais importantes por possuírem o conteúdo mais diretamente relacionada ao conteúdo do documento como um todo.

Uma abordagem como a descrita traz benefícios sobretudo para SRI para a web, onde é comum considerar como relevantes componentes das páginas cujos conteúdos não estejam diretamente relacionados com os assuntos abordados nos documentos, tais como menus, rodapés, etc., e que são tidos como de importância similar a componentes portadores de termos chaves do assunto abordado, tais como títulos e resumos.

Uma vez evidenciada a estrutura de que são compostos os documentos de uma dada coleção, bem como obtidos os pesos inerentes a cada uma das classes estruturais, pode-se adaptar os modelos de RI existentes para que estes tirem proveitos da abordagem apresentada. Em sua dissertação, Oliveira [6] propõe uma extensão do modelo vetorial que inclui as informações de classes e relevâncias.

No modelo vetorial, o cálculo do peso de cada termo perante um documento em uma coleção é obtido por intermédio da combinação de duas evidências: (1) a importância do termo relativo a coleção e (2) a freqüência do termo no documento. Este cálculo não considera as posições de ocorrência e a forma de distribuição dos termos no documento, uma vez que fora concebido para todos os tipos de documentos, tendo eles uma estrutura evidenciada ou não. O modelo proposto defende que uma vez identificada a estrutura de cada documento, bem como o peso de cada classe estrutural dentro de um documento, pode-se se estender o modelo vetorial para obter uma metodologia mais adequada para documentos estruturados. Como proposto pelo trabalho [6], o cálculo da importância do termo no documento é dado pela fórmula:

$$w(t, d) = \sum_{c=1}^{C} tf(t, c) * Peso(c) * idf(t)$$

Sendo que nesta extensão, a freqüência de cada termo é influenciada pelo peso da classe a qual pertence. Os resultados obtidos foram bastante positivos e estão ilustrados junto com os resultados deste artigo na Seção 5.

4 Modelo de Inferência Automática das Classes Estruturais e de suas Relevâncias

Se observarmos este documento como exemplo, perceberemos que este é composto por diversas classes estruturais. Tem-se: título, sub-título, local de execução do trabalho, autores, resumo, títulos das seções e sub-seções e, por último, o conteúdo propriamente dito. Perceberemos ainda que, sem exceção, cada uma destas classes estruturais citadas estão presentes neste documento com formatações de texto diferenciadas. Algumas estão com tamanho maior, outras com fontes diferentes, em negrito, em itálico, enumerado ou qualquer outra combinação destas e outras características de formatação.

Desta forma, fica evidente que um documento estruturado, como este, pode facilmente ser classificado em diversas estruturas onde cada uma contém palavras cuja formatação se assemelham indicando, assim, uma relação direta entre estas palavras. No caso deste documento, o título principal ficaria em uma classe, o resumo em outra, todos os títulos de seções em outra classe e assim por diante.

Uma vez que as classes foram identificadas, basta agora identificar o peso ou a relevância que cada classe apresenta. Está-se querendo inferir, de forma totalmente automática também, que um título, por exemplo, apresenta uma importância, maior do que o resumo. Já este último, apresenta uma importância maior do que o conteúdo dos parágrafos. Assim para todas as classes do documento.

Neste trabalho está sendo proposto a classificação das estruturas de páginas web (ver Seção 4.1) bem como três modelos distintos para o cálculo da relevância destas classes (ver Seção 4.2 e 4.3).

4.1 Estruturas em Documentos Web

HTML, do inglês *Hypertext Markup Language*, é uma linguagem de formatação de textos utilizada para definir a aparência dos documentos que a utilizam. Uma vez que praticamente todas as páginas encontradas na Internet utilizam esta linguagem baseada em formatação, nada mais justo do que utilizar este critério para favorecer as máquinas de busca. Daí a idéia de se utilizar a formatação das páginas como método de inferência das classes estruturais.

Documentos web apresentam um grande número de classes estruturais. No caso de um site de notícias, por exemplo, algumas classes que podem ser identificadas incluem: título da notícia, resumo, conteúdo, figuras, datas, autores (repórteres), outras notícias, menu, espaço publicitário, notas de identificação da empresa e outras opções como imprimir, enviar para outra pessoa, opinar, pesquisar e etc.

O objetivo deste trabalho não é identificar onde está o título, ou onde está o resumo e sim diferenciar cada uma destas estruturas em classes. Em seguida um cálculo de relevância é executado em cada uma. Desta forma não se saberá que uma determinada classe é o título da notícia, mas sim que esta última tem uma relevância muito alta. A Figura 2 ilustra esta idéia.

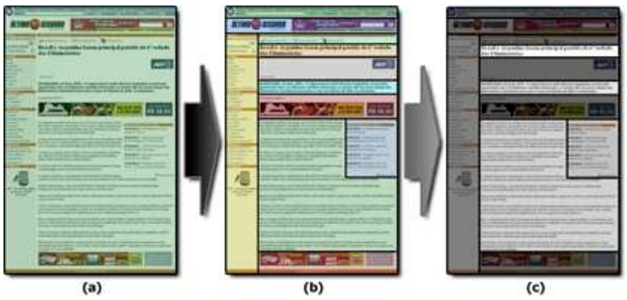

FIGURA 2: *Modelo de Estruturação Automática de Documentos com Base na Formatação*

Inicialmente tem-se um documento da forma como ele é visto por uma máquina de busca tradicional (a). Em seguida o modelo de inferência automática das classes estruturais com base na formatação dos textos, proposta neste trabalho, é aplicado (b). Por último, o cálculo da relevância é realizado, onde as classes mais importantes estão mais claras e as menos relevantes mais escuras (c). Em seu trabalho, Oliveira [6], mostra que a pesquisa em base de documentos como o terceiro mostrado na ilustração (c) apresenta resultados superiores de precisão e revocação do que em bases de documentos não classificados (a).

Para a identificação das classes, foram utilizadas as formatações mais comumente encontradas no HTML: título do documento (<Title>), título de seções (<Hx>), fonte (<Font>), negrito (<B>), itálico (<I>), *link* (<A>), imagem (<IMG>) e os atributos de CSS (Class) e estilo (Style). Qualquer combinação destas formatações encontradas em um documento gera uma nova classe.

4.2 Cálculo da Relevância com Base nas Palavras

Oliveira [6] propôs uma possibilidade de cálculo de relevância baseada na extração da similaridade de um componente estrutural do documento com o restante deste mesmo documento.

Uma fórmula é proposta para o cálculo da relevância de uma classe:

$$Peso(c,d) = \frac{\sum_{t-1}^{T} \frac{NCdt}{NCd}}{T} \qquad (1)$$

onde *c* corresponde à classe, *d* ao documento, *t* a um termo do documento, *T* à quantidade total de termos, *NCdt* à quantidade de classes do documento *d* que possuem o termo *t*, e *NCd* à quantidade total de classes do documento.

A idéia por trás desta equação é que uma classe terá um peso tão maior quanto maior for o número de vezes que os termos que a ele pertence apareçam em outros componentes do documento. Em outras palavras, se os termos pertinentes a uma classe estão também contidas em outras classes do documento, então esta classe terá um alto nível de relevância perante os processos de busca [6].

Ao aplicar esta fórmula nos experimentos realizados neste trabalho, obteve-se bons resultados iniciais. Porém, notou-se que classes que possuíam poucas palavras, mesmo que estas não estivessem tão relacionadas com o documento, obtinham, geralmente, relevâncias altas, parecidas, ou até mesmo superiores que o conteúdo do documento em si. Para o trabalho referenciado, isto não é um problema, visto que a classificação dos documentos foi feita manualmente e dificilmente apareciam classes muito pequenas. Porém, na nossa proposta, muitas classes pequenas aparecem, constituídas de mensagens de *copyrigth*, ou outras poucas palavras que ficaram com formatações diferentes. Este fato prejudicou bastante o cálculo de relevâncias, de forma que uma nova formula está sendo proposta:

$$Peso(c,d) = \frac{\sum_{t-1}^{T} \frac{NCdt+(NtC-1)}{NCd}}{T} \qquad (2)$$

Esta fórmula é parecida com a proposta no trabalho referenciado, porém com a adição do número de vezes que o termo *t* aparece no componente *c* menos um na somatória. Esta modificação basicamente quer dizer que se um determinado termo aparece várias vezes também no componente em questão, então este deve ser beneficiado com alguns pontos de relevância a mais. Os termos pequenos, que geralmente não apresentam termos repetidos, continuam com o cálculo original. Alguns experimentos realizados mostraram que esta última fórmula é capaz de obter resultados melhores do que a fórmula original.

4.3 Cálculo da Relevância com base na Formatação

Este tipo de cálculo está sendo proposto neste trabalho como mais uma possibilidade de cálculo de relevância de uma classe estrutural. Seu princípio é bastante simples: tentar impor uma determinada relevância para uma classe tomando como base a sua característica de formatação. Em outras palavras, não estamos nos importando com as palavras da classe, e sim se elas estão em negrito, itálico, com tamanho grande ou pequeno e etc.

Desta forma, as seguintes classes de formatação foram definidas em ordem decrescente de relevância: título da página, H3, H2, H1, fonte grande, negrito, itálico, sublinhado, texto normal, *link* e, por último, descrição de imagem. Sendo assim, palavras que estão com formatação de título de seção de tamanho 3 (H3) simplesmente serão consideradas mais relevantes do que as palavras com formatações normais.

4.4 Cálculo da Relevância com base em Padrões

A RI realizada em bases de imagens ocorre através do casamento de padrões entre as características de uma determinada imagem e a base em que se está procurando. Este modelo poderia ser trazido como uma proposta de inferência automática da seguinte forma: um título de uma notícia, por exemplo, possui algumas características como tamanho um pouco maior que o restante do texto, algum tipo de destaque (negrito ou itálico), é composta de geralmente poucas palavras e estas palavras costumam se repetir bastante ao longo do texto. Características como essas podem ser geradas para menus, resumos, subtítulos, e etc. Desta forma, uma página web seria analisada e estes padrões seriam procurados nesta página. Cada padrão encontrado constitui uma classe estrutural.

O cálculo de relevância poderia ser feito com um número fixo para cada padrão (e.g., título tem relevância 2.5) multiplicado pelo grau de similaridade entre o padrão e o texto encontrado (e.g., o texto tem 95% de chance de ser um título), resultando na relevância total da estrutura encontrada (e.g., 2.375). Este modelo de inferência automática está apenas como uma proposta neste trabalho e experimentos com tal modelo fica como trabalhos futuros.

4.5 Análise de Complexidade do Modelo

Com a utilização de inferência automática, há um aumento na complexidade da catalogação dos termos. Uma análise mais detalhada da complexidade total ainda não foi feita, mas os passos citados na Seção 5.1 mostram que esta inferência automática é feita em tempo possivelmente polinomial.

5 Implementação e Experimentos Realizados

O objetivo principal deste artigo é a proposta e a implementação de um modelo de inferência automática de classes estruturais e de suas relevâncias em bases de documentos web. O modelo foi visto nas seções anteriores. Nesta seção será discutida a implementação do projeto. Serão discutidos também os experimentos realizados em cima da implementação e, por último, os resultados obtidos pelo modelo proposto.

5.1 Implementação do Modelo Inferência Automática

Basicamente o objetivo desta implementação é receber um documento web como entrada, analisá-lo e gerar como saída um outro documento web dividido em classes estruturais com seus devidos termos (palavras que pertencem à classe) e a indicação da relevância das mesmas.

Desta forma, os principais passos da implementação podem ser citados: (1) fazer um *parse* no documento HTML e gerar uma árvore XML do mesmo. Cada nó da árvore corresponde a um *tag* HTML; (2) percorrer recursivamente a árvore gerada no primeiro passo identificando os *tags* de formatação e seus respectivos conteúdos; (3) adicionar o conteúdo dos nós em suas respectivas classes. Caso não haja uma classe com as características de formatação do conteúdo do nó atual, esta classe deve ser criada. Se já houver a classe, o conteúdo é adicionado a esta última; (4) calcular a relevância de todas as classes geradas. Estes passos devem ser executados para todos os documentos de um determinado diretório (base de documentos).

A implementação do modelo foi realizada utilizando a linguagem C. Para realizar o *parse* no HTML, foi utilizada a biblioteca LIBXML2 do Projeto GNOME. Apesar de ser uma biblioteca de XML, encontra-se, nesta, funções para a manipulação de HTML. Funciona na maioria dos sistemas Unix e Windows.

5.2 Experimentos Realizados

O primeiro tipo de experimento realizado foi baseado na execução da implementação em páginas aleatórias retiradas da Internet e na análise manual das classes e relevâncias inferidas pela implementação.

Um segundo tipo de experimento foi feito em conjunto com o autor do trabalho [6] citado utilizando uma coleção de documentos. Para permitir uma comparação entre a inferência automática, proposta neste trabalho, e a inferência manual, proposta na dissertação citada, foi utilizada a mesma coleção de documentos encontrada nos experimentos deste último trabalho. Tal coleção foi retirada de um site de notícias, o último segundo, do IG, e foram obtidas 13.000 notícias. Foi obtida também um conjunto de consultas que mais regularmente são feitas pelos usuários.

Como medida de comparação, foi-se escolhido traçar o gráfico de precisão e revocação dos seguintes modelos:

(1) modelo vetorial tradicional; (2) modelo vetorial estendido utilizando inferência manual, proposto na dissertação; e os modelos propostos neste trabalho: (3) modelo vetorial estendido com a inferência automática utilizando a fórmula 1 para o cálculo de relevância; (4) utilizando a fórmula 2 e (5) utilizando o cálculo de relevância baseada na formatação.

Antes de ser mostrado os resultados, algumas considerações precisam ser feitas a respeito da coleção utilizada e das fórmulas de relevância propostas. Sobre a coleção utilizada, tem-se que esta não se apresentou nem um pouco favorável ao nosso modelo, visto que este modelo é baseado completamente na formatação dos textos, formatação esta muito pobre nos documentos da coleção utilizada.

Sobre o cálculo das relevâncias utilizando as fórmulas 1 e 2, tem-se que tais fórmulas, consideram uma página isoladamente, sem levar em consideração as outras páginas da coleção. Isto gera um problema: palavras que se repetem muito em todas as páginas (*e.g.,* Último Segundo) deveriam ser desconsideradas mas, ao contrário, estão recebendo as maiores relevâncias. Este problema seria resolvido através de técnicas conhecidas em RI para diminuir a relevância das palavras que se repetem ao longo de muitas páginas. Como estas técnicas não tem muito a ver com classes estruturais nem formatação, estas foram desconsideradas, o que resultou em um ponto negativo nos resultados obtidos.

5.3 Resultados

Os resultados do primeiro tipo de experimento foi um conjunto de documentos gerados e classificados automaticamente pela implementação e não tem como ser apresentado neste artigo. A análise destes documentos mostram que o modelo proposto é capaz de se comportar muito bem para documentos bem formatados e razoavelmente mal para outros com formatação pobre. Verificou-se que em documentos bem formatados, a fórmula de número 2, proposta neste artigo, apresentou os melhores níveis de acerto.

Os resultados obtidos pelo segundo tipo de experimento estão ilustrados na Figura 3. Como pode-se ver, os resultados obtidos para as três fórmulas de cálculo de relevância da inferência automática obtiveram resultados inferiores ao modelo vetorial tradicional e ainda mais inferiores em relação ao modelo vetorial estendido utilizando inferência manual. Tais resultados já eram esperados levando-se em consideração a coleção utilizada e o fato de não se levar em consideração o conteúdo dos outros documentos no cálculo de relevância das classes estruturais. Este último motivo pode ser confirmado observando-se os resultados obtidos pela terceira técnica de cálculo de relevância. Esta técnica, por não considerar as palavras e sim a formatação, não foi prejudicada pela repetição constante de algumas palavras irrelevantes nos documentos. Mostrando, assim, que se esta técnica for utilizada, ainda em uma coleção mal formatada como a atual, resultados muito próximos ao vetorial podem ser conseguidos. Considerando ainda uma base bem formatada, espera-se, então, resultados possivelmente melhores que o modelo tradicional e próximos ao modelo estendido.

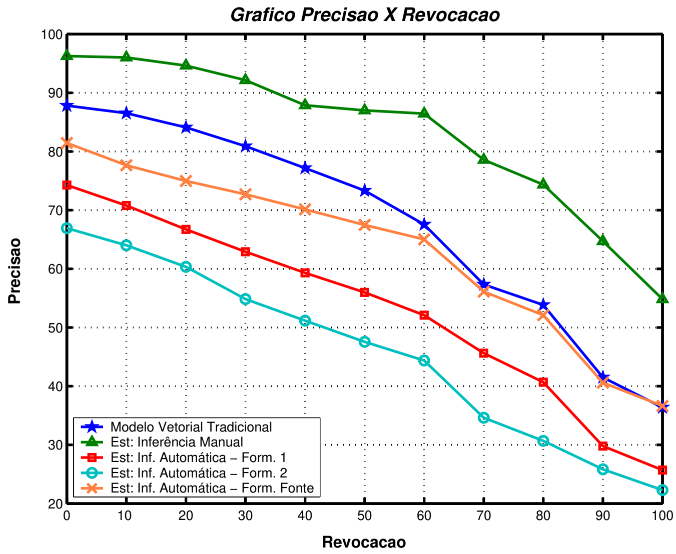

FIGURA 3: *Comparação entre os modelos implementados, o proposto no trabalho citado e o vetorial.*

6 Conclusões e Trabalhos Futuros

O objetivo deste trabalho não é melhorar o modelo estendido, visto que a inferência e o cálculo das relevâncias nos documentos considerados é feito manualmente por pessoas que conhecem os documentos e o assunto e podem, assim, dizer o que é mais importante em uma busca. Algo que torna-se muito difícil de ser feito automaticamente. Porém, espera-se que com as melhorias propostas e com a utilização de bases mais bem formatadas (que são normalmente encontradas na Internet) os resultados podem ser bastante próximas ao modelo vetorial estendido e muito certamente melhores que o modelo vetorial tradicional.

Referências

[1] AL-KHALIFA, Shurug. YU, Cong. JAGADISH, H. V. *"Querying Structured Text in an XML Database"*. Proceedings of ACM SIGMOD, 2003.

[2] BAEZA-YATES, Ricardo. RIBEIRO-NETO, Berthier. *"Modern Information Retrieval"*, Addison-Wesley, 1999.

[3] KENT, A. BERRY, M. M.LUEHRS, L. V. Jr. PERRY, J. W. *"Machine literature searching VIII: Operational criteria for designing information retrieval systems"*. American Documentation, 1955.

[4] KOTSAKIS, Evangelos. *"Structured Information Retrieval in XML documents"*. Proceedings of ACM symposium on Applied computing, 2002.

[5] LIMA, Luciano R. S. *"Categorização Automática de Documentos Médicos"*, Tese de Doutorado, UFMG, 2000.

[6] OLIVEIRA, David B. *"Um Modelo de Recuperação de Informações para Documentos Estruturados"*. Dissertação de Mestrado, UFAM, 2004.

[7] RIJSBERGEN, Van C.J. *"Information Retrieval"*. Butterworths, London, 1979.

[8] SALTON, G. *"Automatic Information Organization and Retrieval"*. MacGraw-Hill, New York, NY, 1968.

[9] VICKERY, B.C. *"Techniques of Information Retrieval"*. Butterworths, London, 1970.