UNIVERSIDADE ESTADUAL DO CEARÁ
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO - MACC

ISMAEL MAGALHÃES PEDROSA ROCHA

ORDENANDO DOCUMENTOS ATRAVÉS DA ANÁLISE DE CONTEXTOS DE BANCO DE DADOS

FORTALEZA – CE
2013

ISMAEL MAGALHÃES PEDROSA ROCHA

ORDENANDO DOCUMENTOS ATRAVÉS DA ANÁLISE DE CONTEXTOS DE BANCO DE DADOS

Dissertação submetida à Coordenação do Curso de Mestrado em Ciência da Computação da Universidade Estadual do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientação: Prof. Dr. Gustavo Augusto Lima de Campos.

FORTALEZA – CE
2013

XXXX Rocha, Ismael Magalhães Pedrosa.
Ordenando Documentos Através da Análise de Contextos de Banco de Dados / Ismael M. P. Rocha, Fortaleza – 2013.
YYp.; il.
Orientador: Prof. Dr. Gustavo Augusto Lima de Campos
Dissertação (Mestrado Acadêmico em Ciência da Computação) – Universidade Estadual do Ceará, Centro de Ciências Científicas.
1. Contexto **2**. Integração BD-RI **3**. Ordenação Relativa. I. Universidade Estadual do Ceará, Centro de Ciências Científicas.

ISMAEL MAGALHÃES PEDROSA ROCHA

ORDENANDO DOCUMENTOS ATRAVÉS DA ANÁLISE DE CONTEXTOS DE BANCO DE DADOS

Dissertação submetida à Coordenação do Curso de Mestrado em Ciência da Computação da Universidade Estadual do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Aprovada em: ___ / ___ / _____ .

BANCA EXAMINADORA

_____

Prof. Dr. Gustavo Augusto Lima de Campos (Orientador)
Universidade Estadual do Ceará – UECE

_____

Prof. Dr. Jorge Luiz de Castro e Silva
Universidade Estadual do Ceará - UECE

_____

Prof. Dr. José Maria da Silva Monteiro Filho
Universidade Federal do Ceará - UFC

AGRADECIMENTOS

Gostaria de agradecer primeiramente a minha esposa Carolina, que é tudo para mim: minha companheira, amiga, confidente e que carrega hoje no seu ventre nossa maior razão de viver, nosso filho Mateus, que mesmo antes do seu nascimento já é muito amado por nós.

Agradeço enormemente aos meus pais, Francisco e Alana, que me criaram com uma sabedoria impressionante, orientando e proporcionando educação de qualidade, saúde e principalmente muito amor, que me fizeram ser o homem que sou hoje.

Agradeço às minhas irmãs, Sabrina e Rebeca, que são muito mais que duas pessoas que possuem o mesmo sangue e sobrenome que eu, são minhas verdadeiras amigas e companheiras de todas as horas.

Quero agradecer também a minha querida e amada avó Cecília, hoje a matriarca da nossa família, poço inesgotável de sabedoria, compaixão e bondade. Sempre preocupada com os filhos e netos, é um exemplo vivo de uma alma iluminada e especial.

Agradeço também ao meu sogro e minha sogra, Haroldo e Mônica, que hoje são como segundos pais para mim, além de serem inseparáveis companheiros de viagens.

Quero também agradecer a todos os meus familiares, meu sobrinho Luquinhas, meu primo Junior e sua esposa Claudiene, ao tio Luiz e a tia Ludmila, aos meus cunhados Luciano e Jota, enfim, a todos aqueles que, de alguma forma, contribuíram para que eu pudesse chegar a esse momento.

Por fim, mas não menos importante, quero agradecer ao meu orientador Gustavo, que apostou no meu potencial desde o inicio, e sempre foi paciente e conselheiro, como poucos que já vi por essa vida.

Quero também registrar meus sinceros agradecimentos ao prof. Marcus Sampaio, que foi co-orientador de boa parte desse trabalho, além de ter sido o responsável pela identificação do tema que serviu de base para essa dissertação.

RESUMO

Estamos vivendo um momento de virada na história da pesquisa em Bancos de Dados, com uma explosão tanto de dados como de cenários de uso. Um problema emergente é como recuperar informação não estruturada (documentos em linguagem natural) que sejam relevantes para uma consulta feita a um BD. Este trabalho apresenta uma nova abordagem para integrar dados estruturados e não estruturados onde, no momento em que o usuário fizer uma consulta a um BD, ele ainda tem a possibilidade de receber informações relevantes provenientes de textos não estruturados, com o diferencial de estes documentos estarem aproximadamente em ordem de relevância para a consulta. Para demonstrar a viabilidade da proposta, é apresentada uma avaliação experimental frente o atual estado da arte.

ABSTRACT

We are at a turning point in the history of the database research, due both to an explosion of data and usage scenarios. A challenging problem is the retrieval of unstructured text documents that are relevant to a structured database query. In this work we present a new approach for integrating structured and unstructured data in which when the user queries a database he still can receive extra relevant information from text sources which is presented approximately in order of relevance to the query. The experimental evaluation of our proposal demonstrated its feasibility.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No mundo atual, onde palavras como “Integração” e “Facilidade” estão em alta, observa-se um crescente número de cenários de gerenciamento de dados envolvendo ao mesmo tempo dados estruturados, ou seja, bases de dados (BD), e dados não estruturados, que são informações armazenadas em artefatos escritos em linguagem natural (documentos técnicos, atas de reunião, etc.), que são tipicamente recuperados pelos mecanismos de recuperação de informação (RI). Nesse contexto, a integração BD-RI tornou-se um dos principais desafios de ambas as áreas [(ABITEBOUL et al., 2005), (WHANG et al., 2005), (WHANG, 2009)].

Os estudos mostram que, até o momento, a integração entre as bases de dados e os sistemas baseados em linguagem natural ainda é um problema em aberto. Uma questão inerente a esta abordagem é a recuperação de documentos que são relevantes para um contexto relacionado a uma consulta de banco de dados. É importante notar que, no domínio de RI, em muitos casos os documentos recuperados por uma máquina de busca não são exibidos em uma ordem consensual de relevância para um determinado contexto relacionado aos termos da pesquisa, o que pode causar desconforto para o usuário. Paradoxalmente, essa questão é em grande parte inexplorada na literatura.

Nesse universo, avaliar de forma eficiente os resultados recuperados pelas máquinas de busca também é de suma importância. As tradicionais métricas utilizadas pela RI não são capazes de distinguir com qualidade a relevância posicional de cada documento, ou seja, se a posição fornecida na lista do ranking do motor de busca é a mais adequada de acordo com a consulta feita pelo usuário.

1.1 Motivação

Para exemplificar que a associação BD-RI é um tema em ascensão podemos tomar como base a última conferência TREC, que foi baseada em avaliações

de relevância binárias de documentos[1]. A *track* que discute sobre filtragem descreve um cenário onde existe a necessidade de informação do usuário, porém há um fluxo grande e contínuo de novos documentos. Para cada documento da coleção (novo ou não), o sistema deve decidir de maneira binária se o mesmo é relevante ou não e se deve ser recuperado.

O Relatório sobre *Claremont Research Database* (AGRAWAL et al., 2008) reforça que estamos num ponto de viragem na história da pesquisa de banco de dados, devido, entre outros aspectos, tanto a uma explosão na quantidade de dados disponíveis assim como nos quase incalculáveis cenários de uso. Um dos fatores para tal é a onipresença de dados estruturados e não estruturados. O relatório identifica pelo menos duas importantes oportunidades:

- Revisitação dos motores de banco de dados: Ampliação da gama útil de aplicabilidade para sistemas multiusos de banco de dados (por exemplo pesquisa de texto e integração de informação);
- Integração entre dados estruturados e não estruturados: O desafio é a descoberta de relações entre os dados estruturados e não estruturados.

Por outro lado, o relatório sobre integração BD-RI (AMER-YAHIA et al., 2008) aponta em seus destaques “Pesquisa Com Contexto” entre as questões quentes e temas emergentes. Aplicações voltadas para banco de dados parecem estar ficando cada vez mais orientadas para o usuário (trazendo a área mais para mais perto da RI, onde os aspectos da consciência dos usuários tem uma longa tradição), ao contrário das clássicas aplicações voltadas para a plataforma de negócios que hoje estão perfeitamente dominadas.

Várias propostas no campo da integração BD-RI se encaixam em uma abordagem que é mais comumente chamada de integração de informação orientada a contexto (MOHANIA, BHIDE, 2008). Nesse cenário, podemos distinguir quatro grandes problemas. O primeiro é que a ligação entre o banco de dados e os documentos é estática (CHAKARAVARTHY et al., 2006), ou seja, sem levar em conta de forma dinâmica as consultas feitas pelo usuário. Seria bem mais relevante

[1] Disponível em: http://trec.nist.gov.

fazer uma ligação com as consultas do usuário, ou seja, definir contextos dinâmicos em vez de contextos genéricos pré-definidos. Em segundo lugar, em (ROY et al., 2005) a hipótese de que palavras-chave cuidadosamente extraídas da consulta ao banco de dados levam a um bom desempenho dos motores de busca, não tem sido verdadeiramente e completamente validada, visto que o próprio o estudo ignora completamente o "lado" relacionado à máquina de busca, concentrando-se exclusivamente no banco de dados. O terceiro problema refere-se a forma como um contexto é gerado e/ou extraído na recuperação de documentos orientada a contexto: essa definição pode ser uma tarefa cara e complexa (CHEN; PAPAKONSTANTINOU, 2011).

Por fim, mas não menos importante, para medir a qualidade das respostas recuperadas pelas máquinas de busca existem hoje inúmeras métricas que permitem avaliar os resultados obtidos. A métrica mais comumente usada para medir os resultados das consultas chama-se *Mean Average Precision* (MAP), que unifica em um mesmo escore os dois elementos clássicos da recuperação de informação: precisão e revocação.

Ao fazer uso de MAP, é possível avaliar se o resultado geral foi satisfatório e a qualidade do conjunto de documentos recuperados foi a melhor possível. Entretanto, o grande problema de MAP é que a mesma não é capaz de distinguir de forma eficiente a relevância posicional de cada documento da lista, uma vez que opera com avaliações binárias [(HOFMANN et al., 2009), (RADLINSKI; CRASWELL, 2010)]. Assim, não é possível identificar se determinado documento que está na posição X não deveria está na $X + 1$, por exemplo, de acordo com a expectativa do usuário. Assim, um dos objetivos desse trabalho é comprovar que a utilização de informações adicionais extraídas de uma consulta a um banco de dados são capazes de melhorar a classificação de documentos coletados de uma máquina de busca com base em um conjunto de termos gerados a partir dessa mesma consulta.

1.2 Objetivos

1.2.1. Objetivos Gerais

Para ressaltar a importância do tema associação BD-RI, é suficiente ter em conta que, já em 2003, as estatísticas mostravam que 85% dos dados digitalizados espalhados pelo mundo eram documentos não estruturados (HRISTIDIS; GRAVANO; PAPAKONSTANTINOU, 2003). Paradoxalmente, os investimentos de pesquisa e tecnologia em gerência de dados residem quase que exclusivamente em dados estruturados, mais especificamente nos Sistemas Gerenciadores de Banco de Dados (SGBD).

Por outro lado, os resultados da pesquisa bibliográfica sobre o tema revelam que muito esforço ainda se justifica na busca por soluções que sejam ao mesmo tempo simples (ou de baixo custo de desenvolvimento) e eficientes (tempos de resposta das consultas).

Dessa forma, propõe-se uma nova solução para o problema de associação BD-RI e uma nova métrica de avaliação que leva em conta a relevância posicional dos documentos em uma lista. A solução para o problema de associação pretendeu atingir os seguintes requisitos:

- Uma articulação dinâmica entre banco de dados e documentos;
- Foco distribuído igualmente entre banco de dados e máquina de busca;
- Utilização de contextos para possibilitar a integração, facilmente inferidos a partir das consultas aos bancos de dados.

A nova métrica de avaliação que, em oposição a estratégia binária de relevância posicional, propõe a avaliação dos resultados das consultas navegacionais de forma não binária. Para isso, deve-se verificar o quão significativa é a presença do resultado da consulta no documento. Mais especificamente, considera-se que quanto mais intensa for a presença de cada elemento do resultado da consulta no documento, maior será o grau de relevância do mesmo, e vice-versa.

A abordagem proposta preenche os requisitos de simplicidade e eficiência. A razão principal para fundamentar a hipótese é que a abordagem é não-intrusiva, isto é,

não requer qualquer modificação nem nos SGBD's e nem nos repositórios de documentos. Outra contribuição importante desse trabalho é comprovar que a utilização de informações adicionais extraídas de uma consulta a um banco de dados são capazes de melhorar a classificação de documentos coletados de uma máquina de busca, com base em um conjunto de termos gerados a partir dessa mesma consulta.

1.2.2. Objetivos Específicos

Os principais objetivos desse trabalham foram desenvolver um novo modelo de *blind feedback* que, apoiado por um contexto de Banco de Dados, fosse capaz de melhorar o ranking dos documentos fornecidos por uma máquina de busca, deixando os mesmos aproximadamente em ordem de relevância, assim como uma nova métrica de avaliação, cujo diferencial é levar em consideração a relevância posicional dos documentos de uma lista.

Para isso, inicialmente foi feito um levantamento das principais estratégias sobre a integração BI-RI hoje existentes no mundo científico. A partir desse ponto, um novo algoritmo foi concebido unindo algumas das melhores práticas, tais como uso de contextos de banco de dados e das estratégias de *tf-idf* para calcular peso de termos, resultando em um novo mecanismo de *blind feedback* capaz de reordenar uma lista de documentos recuperados por um motor de busca seguindo os requisitos anteriormente citados.

Para confirmar as hipóteses, uma avaliação experimental se fez necessária. Assim, para viabilizá-la foi desenvolvido um protótipo ferramental que permitiu os avaliadores especialistas analisarem e julgarem os resultados apresentados. Para montar o espaço amostral dessa avaliação, foram utilizados dois domínios distintos e uma amostra de 25 temas para cada um, totalizando um universo de 50 temas distintos. Por fim, foram selecionados 5 usuários com perfis diferentes para compor o grupo de avaliadores especialistas.

1.2.3. Hipóteses Comprovadas e Metodologia Adotada

As avaliações feitas pelos especialistas tiveram como objetivo comprovar as seguintes hipóteses:

- O algoritmo desenvolvido deveria ser competitivo contra um reconhecido trabalho do Estado-da-Arte;

- O algoritmo desenvolvido deveria ser bem sucedido em melhorar o ranking dos motores de busca;

- Embora MAP possua uma lógica própria que remete à ordenação dos documentos relevantes, deveria-se comprovar que não há nenhuma correlação entre ela e a nova métrica proposta.

Foram escolhidos dois diferentes domínios, sendo um sobre filmes (*Internet Movie Database* - IMDb) e outro sobre bandas musicais (*Open Music Project* - MusicMoz). SCORE foi escolhido como o trabalho do Estado-da-Arte para verificar a primeira hipótese e Google™ foi o motor de busca utilizado para verificar a segunda.

1.3 Estruturação da Dissertação

O presente trabalho encontra-se organizado da seguinte forma: o segundo capítulo apresenta alguns trabalhos relacionados às áreas de integração semântica (ou de informação), relevância graduada (*graded relevance*) e métricas de avaliação, mostrando algumas abordagens mais atualizadas, assim como as ferramentas que auxiliam na execução das tarefas. No terceiro capítulo são apresentadas a abordagem da proposta, notadamente a arquitetura de RANKER, assim como apresenta em detalhes como a nova métrica de avaliação RAS foi concebida. No quarto capítulo são apresentadas as avaliações realizadas com a implementação da abordagem descrita no capítulo anterior, algumas variações feitas e também é realizada uma análise

quantitativa dos resultados obtidos. Em seguida é apresentada a conclusão e os trabalhos futuros. Por fim, são apresentadas as referências bibliográficas utilizadas para a confecção deste trabalho.

2 REVISÃO DE LITERATURA

Aqui será apresentada a revisão da literatura, com foco na identificação do problema alvo que pretende-se atacar e nos principais trabalhos encontrados no atual estado da arte. Este capítulo está dividido em duas seções principais. A seção 2.1 caracteriza melhor os problemas que foram tratados nesta dissertação, e a seção 2.2 apresenta os trabalhos relacionados encontrados na literatura sobre os principais temas.

2.1 Caracterização dos Problemas

Suponha o seguinte cenário fictício: Um usuário envia a seguinte consulta para o banco de dados público *Internet Movie Database* (IMDb)[2]: *"Dê-me os atores e produtores da série de 'TV Murphy Brown: TV Tales'"*.

No Quadro 1 temos nos cabeçalhos das colunas os respectivos atores e produtores desejados, que consistem na resposta estruturada da consulta. Considere então os cinco documentos mais relevantes da Web coletados na mesma ordem em que foram extraídos a partir de palavras-chaves extraídas da consulta e que foram submetidas a uma máquina de busca.

Quadro 1 - Conexão entre documentos e dados estruturados.

Documento		Ator							Produtor	
		Candice Bergen	Robert Pastorelli	Dan Quayle	Joe Regalbuto	Grant Shaud	Charles Kimbrough	Pat Corley	Emily Puk	Jennifer Kahill
1	imdb.pt/title/tt2009259	√	√	√	√	√	√	√		
2	museum.tv/eotvsection.php?entrycode =murphybrown	√	√		√	√	√	√		

[2] Disponível em: http://www.imdb.com/

3	imdb.com/name/nm0000298	√								
4	imdb.pt/title/tt0337685	√	√	√	√	√	√	√	√	√
5	imdb.com/title/tt2009259	√	√	√	√	√	√	√		

Espera-se que os cinco documentos recuperados sejam relevantes, ainda que em graus variados de relevância em relação a consulta realizada ao BD. Entretanto, será que a ordem em que os mesmos foram recuperados pela máquina de busca está correta em termos de graus de relevância? Para responder essa pergunta, considerou-se que o grau de relevância de um documento em relação a uma consulta deveria considerar a presença dos resultados da consulta no mesmo. Mais especificamente, considerou-se que quanto mais intensa for a presença do resultado da consulta no documento, maior será o seu grau de relevância, e que quanto menos intensa, menor o grau.

Assim, de acordo com a noção de grau de relevância informalmente definida no parágrafo acima, realizando-se uma análise dos resultados na Tabela 1, pode-se confirmar que os documentos retornados têm realmente diferentes graus de relevância, entretanto não foram ordenados corretamente. Nos extremos, temos os documentos das linhas 3 e 4 que, respectivamente, obtiveram o pior e o melhor graus de relevância. O documento na linha 3 é uma biografia da atriz *Candice Bergen*, e só muito indiretamente faz referência às várias séries *Murphy Brown* em que ela participou, incluindo a série da consulta. O documento na linha 4, apesar de supostamente ser somente o quarto mais importante, é o mais relevante, pois contém todos os atores e produtores da série *Murphy Brown*. De fato, o documento é intitulado *"Murphy Brown: TV Tales (2002) (TV)"*, contendo todo o elenco e grupo de uma das séries de TV.

Com relação aos outros três documentos descritos nas linhas 1, 2 e 5, eles parecem estar entre os dois extremos acima citados: na verdade, eles se referem a outra série de *Murphy Brown* com produtores diferentes, embora ainda com muitos dos mesmos atores. Dessa forma, é altamente desejável que, no momento em que esses documentos forem apresentados ao usuário, que pelo menos o documento da linha 4 seja classificado no topo enquanto o documento da linha 3 deve aparecer como o último da lista.

Dessa forma, este trabalho visou resolver dois problemas. No primeiro, sobre a integração BD-RI, a premissa subjacente é que o conjunto-resposta de uma consulta a um banco de dados deve ser capaz de ajudar um motor de busca a recuperar documentos aproximadamente em ordem de relevância. O problema é definido mais especificamente a seguir: considerando as respostas estruturadas de consultas de usuários a banco de dados, uma sequência de documentos não estruturados na Web classificada por uma máquina de busca a partir da submissão destas respostas, como melhorar a classificação dos documentos coletados pela máquina em termos de graus de relevância?

O problema enunciado deu origem ao segundo problema causado pela ausência na literatura de uma métrica capaz de avaliar de maneira não binária o grau de relevância de um documento em uma coleção considerando as palavras-chave que foram submetidas a uma máquina de busca. Como medir adequadamente a qualidade dos resultados recebidos das máquinas de busca? Assim, o segundo problema esteve na formulação de uma métrica para o grau de relevância satisfazendo as seguintes propriedades enunciadas anteriormente: quanto mais intensa for a presença do resultado da consulta no documento, maior será o seu grau de relevância, e vice-versa.

2.2 Estado da Arte

Considerando os problemas definidos na seção anterior e as abordagens concebidas nesta dissertação, o estado da arte pesquisado pode ser categorizado em três grandes temas: métricas de avaliação binárias, relevância graduada (*graded relevance*) e integração semântica (ou de informação). Assim, esta seção foi dividida em duas subseções correspondentes aos temas.

2.2.2 Relevância Graduada

A natureza distribuída das informações disponíveis na Internet levou à academia a buscar maneiras eficientes de executar consultas sobre uma grande coleção de documentos, através do uso de motores de busca. Entretanto, o conjunto de documentos relevantes para uma consulta pode facilmente ter milhares de itens. Nesse cenário, a montagem de uma ordenação eficaz dos resultados passou a ser uma tarefa crucial. Por exemplo, a estrutura de links, comum nos documentos HTML, permite estimar quais são as páginas mais relevantes, ou seja, partindo-se da premissa que quanto maior o número de links que apontam para certa página (links esses vindos também de páginas igualmente relevantes), mais importante é a página.

Os primeiros algoritmos para ordenação dos resultados das consultas submetidas a um motor de busca eram dependentes da consulta e realizavam uma análise sobre o conteúdo das páginas para montar um ranking. Basicamente era verificado se as palavras-chave da busca estavam no título da mesma, ou estavam nos metadados de descrição e quantas ocorrências existiam dessas palavras-chave no conteúdo. De posse desses valores, era atribuído um escore a cada página que satisfazia os critérios da consulta e se montava a ordenação com base nos valores obtidos.

Em 1996 surgiram os algoritmos que consideravam a relevância de um documento. Em geral, os escores eram computados através de um modelo de análise de links. Esse modelo faz uma análise de um grafo, cujos vértices são as páginas e os links são as arestas. Considerando que um link da página A para B é um “voto de popularidade” para a página B, quanto maior o número de links (vindo de páginas também relevantes) que apontam para certa página, maiores as chances de esta página ter um escore alto e consequentemente ser relevante para a consulta.

Os autores de (JÄRVELIN; KEKÄLÄINEN, 2002) evitam a avaliação binária (relevante ou não) para calcular a relevância dos documentos. Em vez disso, eles propõem a noção de ganho acumulado das pontuações de relevância dos documentos recuperados ao longo da lista de resultados classificados. Para isso, assumem categorias de relevância no intervalo entre 0 e 3 (valor 3 denota valor alto, 0 nenhum valor). Assim, um documento está na posição certa se os outros documentos que o precedem também estão. Até onde as pesquisas avançaram, trata-se do primeiro modelo a descartar a avaliação binária da relevância dos documentos.

2.2.1 Integração Semântica

Os autores de (DOAN; HALEVY, 2005) discorrem basicamente sobre o conceito de integração semântica, ou seja, relacionar e/ou unir bases de dados heterogêneas, e muitas vezes geograficamente dispersas, de forma que estas possam prover as suas informações de maneira unificada e sem duplicações. De acordo com o trabalho, os principais desafios da integração semântica, que também fazem parte do problema da integração BD-RI, são:

- Como decidir qual correspondência um determinado elemento de uma base *S* deve ter numa base *T* para um mesmo conceito do mundo real, visto que na maioria dos casos, essa tarefa é subjetiva e carente de fontes. Por exemplo, elementos com o mesmo nome em bases diferentes podem referenciar conceitos diferentes no mundo real, assim como o contrário (entidades com nomes diferentes correspondendo ao mesmo conceito).

- A documentação que normalmente é utilizada para auxiliar no processo (fontes de dados, *schemas*, etc.) é muitas vezes incompleta.

- É um processo extremamente custoso, pois toda associação necessita de uma verificação no restante dos elementos para se ter a confirmação de que trata-se da melhor escolha.

- As principais técnicas para a solução do problema de integração semântica são semiautomáticas, ou seja, necessitam do auxilio de um especialista, e se dividem em dois grupos: as soluções baseadas em regras e as baseadas em aprendizado.

Resumindo, o trabalho faz um *survey* sobre integração semântica, falando desde a integração de *schemas* heterogêneos até as principais arquiteturas e técnicas descritas na literatura para resolver o problema.

O problema central de (HRISTIDIS et. al., 2003) é encontrar uma forma de efetuar pesquisas em banco de dados por similaridade (conceito-base da área de recuperação da informação e que já estão integrados na maioria dos SGBDs atuais) sem a necessidade de se especificar em que coluna da tabela procurar e ainda assim obter resultados satisfatórios a nível de atributos. Mais ainda, os autores citam que a proposta recupera somente os TOP-K resultados mais relevantes (utilizando os conceitos de ranking), de forma a evitar processamento desnecessário. O ponto de partida dos autores foram os trabalhos DBXplorer e DISCOVER (inclusive serviram de base para a arquitetura do sistema proposto no respectivo trabalho) cujo foco era adicionar pesquisa por palavras-chave em banco de dados relacionais.

Como contribuições, os autores citam que a estratégia adotada permitiu recuperar resultados relevantes mesmo quando os documentos não continham todas as palavras-chave da consulta. Além disso, citam que foi possível recuperar somente os k resultados mais relevantes, e informam que as pesquisas em máquinas de busca apontam que na maioria dos casos os resultados procurados estão entre os 10 e 20 primeiros do conjunto-resposta. Por fim, é apresentado um experimento que atesta as vantagens de performance do modelo proposto.

Já EROCS (CHAKARAVARTHY et al., 2006) é um modelo para promover a conexão de um determinado documento com dados relacionais relevantes, ou seja, é proposto um sistema para relacionar bases de dados estruturadas com documentos (em geral textos em linguagem natural), com o diferencial de identificar e apontar os segmentos do documento onde se encontram as conexões com as chamadas entidades do BD. Para tornar a proposta escalável, é apresentada uma estratégia de cache dos dados, de forma a manter a quantidade de informação recuperada da base de dados no mínimo possível. Funciona da seguinte maneira: Um especialista define um contexto de banco de dados como sendo uma tabela pivô que esteja relacionada direta ou indiretamente a outras tabelas através de chaves estrangeiras (denominadas de tabelas derivadas). A conexão contexto-documento poderá ocorrer mesmo em casos onde o termo de um documento não aparece diretamente na tabela pivô, mas somente em uma das tabelas derivadas. É apresentado um estudo experimental que, de acordo com os autores, procura atestar a eficácia da

abordagem proposta em relação as existentes, tanto no que diz respeito a precisão dos resultados quanto execução e sobrecarga de espaço.

(ROY et al., 2005) descreve SCORE, um modelo para a obtenção de um conjunto de palavras-chave a ser submetido para um mecanismo de pesquisa. O conjunto é obtido como resposta de uma consulta feita a um banco de dados e informações adicionais relacionadas ao respectivo banco de dados. A aquisição das palavras-chave é feita de maneira automatizada e transparente para o usuário, e é disparada assim que é identificado o conjunto resposta da consulta. São consideradas palavras-chave candidatas:

- Um valor de uma coluna de tabela explicitamente mencionada na consulta;
- Um valor de uma coluna cuja tabela esteja nas vizinhanças de uma tabela da consulta[3].

Cada termo do conjunto-resposta da consulta deve ter um peso que deve estar acima de um patamar mínimo pré-definido, e que por sua vez é calculado com base no paradigma *tf-idf* da RI. A lista de palavras-chave é ordenada por ordem decrescente de seus pesos. O tamanho da lista é fixado por um parâmetro. Os contextos (mesmo conceito aplicado em EROCS) são gerados dinamicamente, em tempo de processamento das consultas. Uma vez submetida uma consulta SQL, SCORE segue as seguintes etapas:

1. A consulta é processada;
2. O contexto da consulta é gerado;
3. O contexto gera um conjunto de termos que são submetidos a uma máquina de busca;
4. Os resultados da consulta e da máquina de busca são apresentados.

A principal crítica a SCORE diz respeito ao seu estudo experimental: a hipótese de que o conjunto de palavras-chave extraídos da consulta ao banco de dados conduz a um bom desempenho de um motor de busca não foi validado. O uso da

[3] Esteja direta ou indiretamente ligada a uma tabela da consulta por relacionamentos de integridade referencial.

métrica *Mean Reciprocal Rank* (MRR) não parece suficiente para avaliar a qualidade dos resultados obtidos. Na verdade, o estudo parece ignorar o “lado” RI do problema, concentrando-se exclusivamente em “lado” do banco de dados, supondo que as palavras-chaves escolhidas são as melhores possíveis e inevitavelmente trarão os documentos não estruturados corretos e na ordem de relevância esperada.

(FAGIN et al., 2010) apresentam um *framework* que está encapsulado em algo que os autores chamam de sistema de pesquisa de banco de dados (*Search Database System - SDBS*). De acordo com os autores, um SDBS compreende um esquema, uma instância de banco de dados e uma gramática. As regras gramaticais são elaboradas manualmente por um especialista, através do esquema anteriormente citado. Consultas submetidas a uma máquina de busca são previamente interpretadas com a ajuda da gramática e da instância de banco de dados. O resultado é uma nova consulta que é realmente submetida na forma de palavras-chave para o motor de busca. O raciocínio é então o de levar a lógica de negócio incorporada na base de dados para refinar as palavras-chave. Infelizmente, a fase de trabalho ainda é preliminar, ou puramente teórico, visto que nenhum experimento para provar a eficácia da proposta é mostrado. O trabalho (FAGIN et al., 2011) é um refinamento das ideias expressas em (FAGIN et al., 2010), mais da mesma forma ainda sem uma avaliação experimental.

(CHEN; PAPAKONSTANTINOU, 2011) focam a classificação (ou ranking) sensível ao contexto para recuperação de documentos. Um contexto é definido como sendo um subconjunto de documentos, e é especificado por termos de consultas fornecidas por domínios específicos dos usuários. O ranking do resultado da consulta é calculado com base em estatísticas dos termos coletados a partir desses contextos. Além disso, visões materializadas são usadas com o intuito de melhorar a eficiência da consulta. Vale salientar que, tanto no exemplo como nas experiências, a coleção de documentos dispõe de uma Ontologia, ou seja, um modelo de dados que representa um conjunto de conceitos dentro de um domínio juntamente com os relacionamentos entre estes (GRUBER, 1993), o que realmente facilita a extração de contextos.

2.2.1 Métricas de Avaliação

Como já dito, pretende-se comprovar que a utilização de dados referentes a uma consulta a um banco de dados são capazes de melhorar um ranking de documentos coletados de um motor de busca. Assim, vale entender os conceitos implícitos no algoritmo que tradicionalmente classifica os resultados das consultas feitas a uma máquina de busca, o *PageRank* (PAGE et al., 1999), e nas métricas de avaliação dos resultados produzidos.

O *PageRank*, por definição, é um algoritmo que calcula um número que mede a reputação de uma página ou documento de uma coleção, e consequentemente, quanto maior seu valor, maior é a reputação desse documento. O algoritmo tem como principal objetivo simular o comportamento de um usuário ao navegar na Internet. Em outras palavras, o número do *PageRank* de uma página diz respeito a quão fácil ou difícil seria encontrar esta página partindo de uma URL randômica e, seguindo os links, se tentasse chegar a mesma. Quanto maior o valor resultante do algoritmo, mais fácil seria encontrar a página procurada.

Para medir a qualidade dos resultados gerados por *PageRank* e demais algoritmos de ordenação, foram criadas, ao longo dos anos, várias métricas de avaliação como, por exemplo, a precisão e a revocação. A primeira delas mede o percentual de documentos relevantes no universo de documentos retornados pelo motor de busca (MANNING; RAGHAVAN; SCHTZE, 2008). A precisão é dada pela seguinte fórmula:

$$precisão = \frac{DocumentosRelevantes \cap DocumentosRecuperados}{DocumentosRecuperados} \qquad (2.1)$$

A fórmula indica que a métrica busca a cardinalidade do conjunto interseção entre os documentos relevantes e a lista de documentos recuperados,

dividido pelo total de documentos recuperados. Para identificar tais relevantes, a lista de documentos recuperados deve ser apresentada a uma especialista que indicará quais são os que têm valor para a consulta.

A precisão considera todos os documentos retornados pelo motor de busca. Entretanto, uma adaptação bastante utilizada é considerar um ponto de corte nos documentos retornados, devido ao número elevado de documentos retornados. Desse corte, surgem as métricas denominadas P@N (ou TOP-K@N), onde N é o ponto de corte. Por exemplo, a métrica P@10 considera, para cálculo da precisão, somente os dez primeiros documentos retornados. Nesse caso, se entre esses 10 encontram-se 3 relevantes, a P@10 é igual a $\frac{3}{10}$, ou seja, 0,3 ou ainda 30%.

Já a revocação mede qual o percentual de documentos relevantes recuperados por uma pesquisa perante o número total de documentos relevantes na coleção. A revocação é calculada pela seguinte fórmula:

$$revocação = \frac{DocumentosRelevantes \cap DocumentosRecuperados}{DocumentosRelevantes} \quad (2.2)$$

A revocação é utilizada em conjunto com a precisão para calcular a *Mean Average Precision* (MAP). A métrica MAP unifica em um mesmo escore esse dois valores clássicos da recuperação de informação (HOFMANN et al., 2009), e é descrita da seguinte maneira:

$$MAP = \frac{\sum_{q \in Q} AP(q)}{|Q|} \quad (2.3)$$

Na fórmula, *Q* é o número de consultas avaliadas e *AP(q)* é a precisão média da consulta *q*. A métrica calcula a média das precisões médias (*average*

precision) de cada consulta entre as consultas que estão sendo avaliadas. Por sua vez, a precisão média de uma consulta é a média das precisões calculadas após cada documento relevante recuperado.

Entretanto, a principal deficiência dessa e das demais métricas tradicionais da RI se apresenta quando se necessita calcular a importância relativa dos documentos em cada ranking, pois as mesmas não são capazes de distinguir bem relevância posicional, uma vez que operam somente com avaliações binárias (RADLINSKI; CRASWELL, 2010). Por exemplo, considere o Quadro 2, descrito abaixo, que reflete as mesmas informações contidas no Quadro 1 mas agora acrescido de uma nova coluna, que apresenta a avaliação de relevância do especialista para cada documento da consulta ao IMDb:

Quadro 2 - Apresentação da métrica MAP.

Documento		Ator							Produtor		Especialista
		Candice Bergen	Robert Pastorelli	Dan Quayle	Joe Regalbuto	Grant Shaud	Charles Kimbrough	Pat Corley	Emily Puk	Jennifer Kahill	
1	imdb.pt/title/tt2009259	√	√	√	√	√	√	√			√
2	museum.tv/eotvsection.php?entrycode=murphybrown	√	√		√	√	√	√			√
3	imdb.com/name/nm0000298	√									√
4	imdb.pt/title/tt0337685	√	√	√	√	√	√	√	√	√	√
5	imdb.com/title/tt2009259	√	√	√	√	√	√	√			√

Como todos os documentos contêm pelo menos um dos elementos da resposta do BD, todos são considerados relevantes pelo especialista. Dessa forma, a métrica terá valor máximo igual a 1 (a lista de documentos relevantes é igual a de documentos recuperados). Aqui a deficiência de MAP fica exposta, pois apesar de todos os documentos serem relevantes, existe uma graduação relativa que deveria ser levada em consideração. Como descrito na primeira seção deste capítulo, o documento listado na posição 4 é o mais aderente ao contexto da consulta, portanto deveria estar na primeira posição da lista. A métrica MAP, por trabalhar com o

conceito binário de relevância, não consegue diferenciar essa relação, não permitindo portanto avaliar com qualidade tais resultados.

3 MATERIAIS E MÉTODOS

Este capítulo descreve as abordagens para resolver os dois problemas enunciados na seção 2.1. A abordagem para o primeiro foi denominada RANKER. Primeiramente, o capítulo foca na arquitetura do algoritmo RANKER, um sistema composto de dois pilares principais: Integração BD-RI com base na noção de contexto (extraído da consulta a um banco de dados); esse contexto orienta a ligação entre os dados estruturados recebidos como resposta a uma consulta de banco de dados, e informações não estruturadas recuperadas por uma máquina de busca. A abordagem para o segundo problema foi denominada RAS. O capítulo também detalha a formulação de RAS, que tem como diferencial o fato de levar em consideração a relevância posicional de documentos.

O capítulo foi dividido em seis seções principais. A seção 3.1 apresenta uma visão geral da arquitetura de RANKER e de seu funcionamento. As seções 3.2, 3.3 e 3.4 especificam os principais passos envolvidos no funcionamento do algoritmo. Em seguida, na seção 3.5, mostraremos como foi concebida a nova métrica de avaliação RAS e, por fim, a seção 3.6 apresenta os detalhes da amostragem utilizada nos experimentos.

3.1 Arquitetura de RANKER

A Figura 1 ilustra uma visão da arquitetura do sistema RANKER. Seu funcionamento se baseia em um processo composto de três passos:

P1. Passo BD – O usuário faz uma consulta a um Banco de Dados;

P2. Passo RI – O sistema faz uma consulta por palavra-chave a uma máquina de busca;

P3. Passo *Blind Feedback* – Sistema reordena a lista de documentos recuperados deixando-os aderente ao contexto da consulta ao BD.

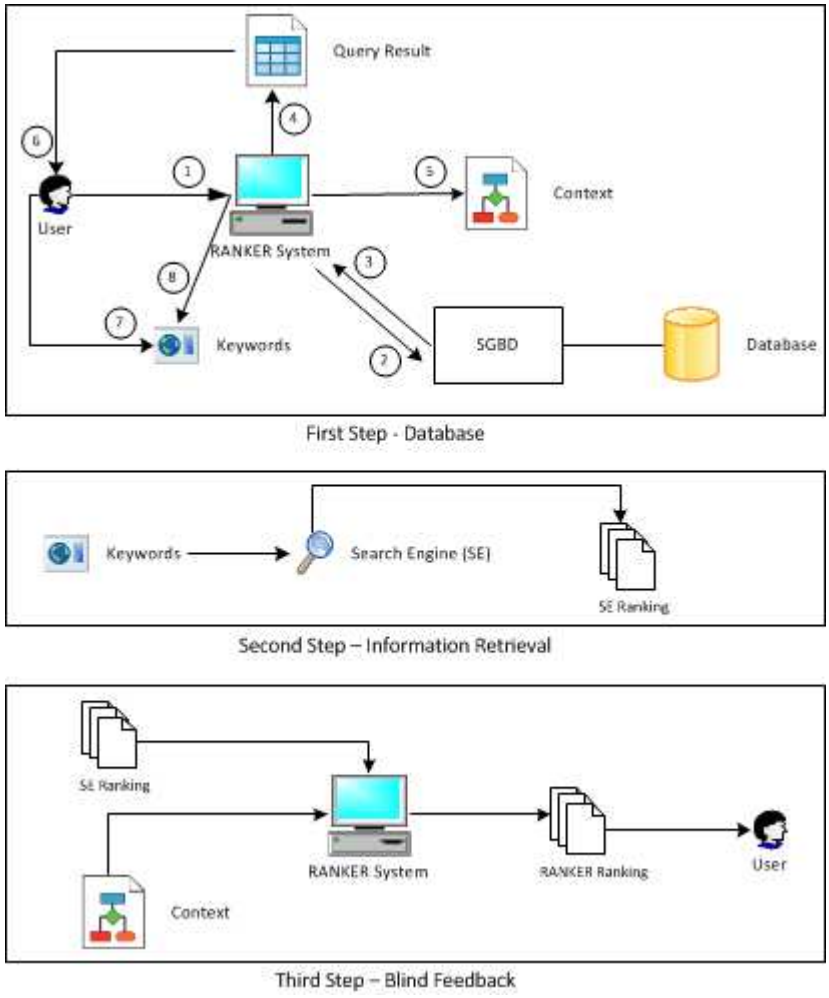

Figura 1 - Arquitetura de RANKER.

No primeiro passo (P1) o usuário deve informar o que deseja buscar no Banco de Dados, ou seja, a consulta em questão. O sistema, ao executar a consulta, utilizará o resultado para inferir um contexto e definir as palavras-chave que serão submetidas à máquina de busca. No passo (P2), os termos obtidos em P1 são enviados ao motor de busca, o resultado é coletado mas não é apresentado ao usuário. Finalmente, no terceiro passo (P3), com a ajuda do contexto inferido em P1, a lista de documentos é reordenada e apresentada ao usuário.

As seções a seguir detalharam passo a passo de cada um dos três passos do processo.

3.2 Passo BD

Na etapa de consulta ao banco de dados o usuário envia normalmente uma consulta SQL para um BD gerenciado por um Sistema de Gerenciamento de Banco de Dados (SGBD). Por exemplo, suponha que o usuário envia para o SGBD que gerencia o banco de dados do IMDb[4] a seguinte consulta: "*Dê-me os atores e produtores da série de TV 'Murphy Brown: TV Tales' "*. A formalização da consulta SQL é mostrado na Figura 2:

```
Select a.(name || surname) As Actor, pb.(producerN || producerS) As Producer
From Movies m, Actor a, Actor2Movie am, ProducedBy pb
Where m.title = 'Murphy Brown: TV Tales' and m.id=am.movie And am.actor=a.id And
      m.id=pb.movie
```

Figura 2 - Consulta SQL.

A lista de atores apresentados na resposta da consulta são {'Candice Bergen', 'Robert Pastorelli', 'Dan Quayle', 'Joe Regalbuto', 'Grant Shaud', 'Charles Kimbrough' e 'Pat Corley'}. Os produtores do conjunto-resposta são {'Emily Puk' e 'Jennifer Kahill'}.

O resultado da consulta é a base para a geração do contexto. A Figura 2 mostra que a tabela *Actor* está associada às tabelas *Movies, Actor2Movie* e *ProducedBy* através de chaves estrangeiras. Assim, considera-se que *Actor* é a tabela pivô e as demais são as derivadas. O contexto é elaborado através de uma consulta que é feita ao DB que recupera todas as colunas, inclusive aquelas que não estejam presentes no conjunto-resposta dessa consulta. Assim, apesar de o usuário desejar recuperar somente as colunas *name* e *surname* de *Actor* e *producerN* e *producerS* de

[4] O script utilizado foi extraído de: http://www.dis.uniroma1.it/%7Edegiacom/didattica/semingsoft/seminari-studenti/08-09-09%20-%20SIS%20-%20Valerio%20Del%20Grande%20-%20Junio%20Valerio%20Franchi/DB/

Producer, o contexto a ser formado através dessa consulta recuperará todas as colunas que estejam *Actor* e nas tabelas derivadas.

A partir dai, esse resultado é apresentado ao usuário para que esse indique os termos que, em sua opinião, melhor referenciam a consulta. Esses termos indicados, juntamente com os literais da consulta, formarão o conjunto de palavras-chave que deverão ser submetidos para a máquina de busca. Por exemplo, supondo-se que o conjunto de palavras-chave escolhido pelo usuário foi {Murphy Brown TV Tales Candice Bergen}, foram selecionadas quatro palavras-chave do conjunto que compõem o literal da consulta SQL e uma palavra-chave da resposta à consulta. Estas são as palavras que serão submetidas à máquina de busca.

Vale ressaltar, o usuário pode selecionar toda a resposta da consulta se quiser, mas dificilmente essa estratégia funcionará, pois as máquinas de busca costumam limitar o número de palavras-chave. Na prática, a maioria das consultas de palavras-chave é realizada com no máximo de três palavras-chave[5]. A máquina de busca retorna um número predefinido de documentos ordenados, cuja ordenação dos documentos é transparente para o usuário.

3.3 Passo RI

Essa etapa se inicia após a seleção das palavras-chave que serão submetidas à máquina de busca.

O objetivo é recuperar os documentos idealmente mais relevantes para a consulta ao BD. Por exemplo, supondo-se que o conjunto de palavras-chave escolhido pelo usuário submetido à máquina de busca Google, coincidentemente, resulta nos 5 documentos melhores ordenados na Tabela 1[6], e conforme discutido no início do Capítulo 2, deixa a desejar em termos da ordem em que documentos relevantes apareceram no resultado. No exemplo Google não foi sensível à relevância do

[5] Afirmação extraída de: http://www.experian.com/hitwise/press-release-bing-powered-share-of-searches-at-29-percent.html
[6] Consulta feita em 21/11/2011.

documento considerado mais importante, ou seja, o imdb.pt/title/tt0337685, que aparece apenas na quarta posição.

Analisando estes tipos de resultados, percebe-se a real necessidade do terceiro passo do algoritmo RANKER, denominado *blind feedback*. Além disso, muitas vezes a falta de paciência do usuário faz com que somente os dois primeiros documentos do conjunto-resposta sejam examinados. Por isso, a etapa do processamento de RANKER evita que a ordenação retornada pela máquina de busca seja diretamente exibida para o usuário, provocando uma reordenação dos resultados obtidos que seja, por exemplo, mais relevante que os obtidos no Quadro 1.

3.4 Passo *Blind Feedback*

É nessa etapa que realmente entra em cena o algoritmo RANKER, cuja função é fazer a ligação entre os dados estruturados retornados pelo BD e a lista de documentos retornados no passo da consulta por palavras-chave. Para isso, é usado o contexto gerado dinamicamente no passo anterior, o qual é importante para restringir o espaço de busca de documentos e trazer mais confiabilidade ao mesmo.

Assim, o mecanismo de *blind feedback* verifica cada documento retornado pela máquina de busca procurando pela presença dos termos do contexto no mesmo, ponderando os termos de acordo com critérios apresentados a seguir. O peso de um documento é então obtido como resultado de uma operação que agrega os pesos dos termos do contexto encontrados nesse documento. A lista original é então reordenada em ordem decrescente de pesos de cada documento, resultando em uma nova lista ordenada, que será então apresentada para o usuário. As subseções a seguir formalizam o modelo.

3.4.1. Modelo Formal do Contexto

Considere uma consulta conjuntiva SQL *Q* sobre um banco de dados, cujo conjunto-resposta é dada por *R(Q)*. Adota-se a premissa de que os literais do corpo da consulta também estão em *R(Q)*. Conforme necessário, pseudo-colunas são criadas, da seguinte maneira: considere uma cláusula WHERE com o literal *Li* em *Q*, onde *i* varia no intervalo de 1 a *n*, que é o total de linhas do resultado. Assim, o contexto de *Q*, chamado de *C(Q)*, é dado por $C(Q) = \cup_{i=1}^{n} L_i$. Note que, devido a inclusão das pseudo-colunas com os literais, na maioria dos casos $R(Q) \neq C(Q)$. Por exemplo, o contexto da consulta SQL anteriormente citada é: Lista de Atores ∪ Lista de Produtores ∪ {'Murphy Brown: TV Tales '}.

Se compararmos com a formalização de contexto apresentado por SCORE (ROY et al., 2005), o modelo aqui proposto é totalmente diferente, pois não descarta os termos com menor peso do resultado, como SCORE faz. O modelo de RANKER se assemelha mais com o apresentado em EROCS (CHAKARAVARTHY et al., 2006), com o diferencial de adicionar ao mesmo os literais da consulta, como forma de garantir aderência do contexto gerado a consulta do Banco de Dados. Além disso, outra grande diferença entre as duas abordagens é que aqui as entidades de interesse do banco de dados são definidas dinamicamente, enquanto em EROCS as mesmas são definidas estaticamente.

3.4.2. Peso de um Termo

Considere *T* como sendo uma tabela pertencente a base de dados. Seja *t* um termo para denotar uma instância de uma coluna *Col* de *C(Q)*, temos que cada $t \in C(Q)$ recebe um peso *w(t)*, calculado através da seguinte fórmula:

$$w(t) = log\left(\frac{|T|}{1 + rf(t)}\right) \tag{3.1}$$

$|T|$ é o número de linhas da tabela e *rf(t)* é definido como sendo a frequência com que *t* aparece na base. A expressão $((1 + |T|) / (1 + rf(t))$ mede o inverso dessa frequência (ou a raridade) de *t* no BD. No extremo, quando *rf(t)* tende a 0, *w(t)* tende ao valor máximo.

A definição para o peso é, em parte, inspirada na estratégia *tf-idf* existente na literatura do paradigma da RI (BAEZA-YATES; RIBEIRO-NETO, 1999), a qual calcula a relevância de um termo na consulta considerando que quanto maior a presença do termo em um documento, com relativamente poucas referências a ele no restante da coleção, maior é a relevância do termo para esse documento. Por exemplo, se tomarmos o termo '*Candice Bergen*' recuperado de *R(Q)*, que aparece em somente 131 das 1.362.722 linhas da tabela *Actor*, facilmente podemos entender que trata-se de um termo raro e potencialmente importante para *Q*, e que terá um alto valor de *w(t)*. Os pesos dos termos escolhidos pelo usuário '*Murphy Brown: TV Tales*' e '*Candice Bergen*', citados anteriormente, são respectivamente 4,778 ($|T| \rightarrow |IMDb.Movies| = 599.864$ e $rf = 9$) e 4,014 ($|T| \rightarrow |IMDb.Actor| = 1.362.722$ e $rf = 131$). Os termos que não são escolhidos pelo usuário para compor as palavras-chave também terão seus pesos calculados e serão utilizados nas análises dos documentos.

Curiosamente, em alguns casos ocorre de os termos com os maiores pesos (ou os termos mais raros do BD) serem elementos pouco interessantes para o usuário, mas que serão essenciais para distinguir os documentos mais aderentes ao contexto em relação ao restante da coleção. No nosso exemplo, '*Emily Puk*' e '*Jennifer Cahill*' ambos com 5,033, tem os maiores pesos. Apesar deles não terem sido selecionados como palavras-chave pelo usuário, são importantíssimos para o sistema.

Vale ressaltar que, na concepção de RANKER, os termos podem ser tantos palavras simples como termos compostos. Essa situação é transparente para o algoritmo, deixando-o mais robusto, pois permite adequar o cálculo dos pesos de acordo com o BD utilizado. Por exemplo, no IMDb é mais conveniente buscar pelos nomes completos dos atores e produtores, portanto deve-se configurar um parâmetro no algoritmo para calcular os pesos sem separá-los. Em outros casos, como, por exemplo, consultas a um BD de gêneros musicais, pode-se querer recuperar de uma consulta tanto documentos com os termos individuais "Pop" e "Rock" como o termo composto "Pop Rock".

3.4.3. Cálculo do *Score* de um Documento

Ranker considera o documento *D* como sendo uma lista de termos. Seja *t-set(D)* o conjunto dos termos que aparecem simultaneamente em *D* e *C(Q)*. A relação entre eles é definida da seguinte forma:

$$\mathrm{score}(\mathrm{D} \leftrightarrow \mathrm{C}(\mathrm{Q})) = \sum_{\mathrm{t} \in \mathrm{t-set(D)}} \mathrm{tf}(\mathrm{t}, \mathrm{D}) \times \mathrm{w}(\mathrm{t}) \qquad (3.2)$$

Assim, o escore de um documento é dado pelo somatório de todos os termos do documento que pertencem ao contexto, onde para cada termo é feito o produto da frequência do termo *t* no documento *D* [*tf(t, D)*] pelo peso do termo [*w(t)*].

É importante reforçar que RANKER não está limitado aos termos escolhidos pelo usuário para compor as palavras-chave, uma vez que, como já foi dito, ele calcula o peso de todos os termos existentes no contexto, ampliando assim sua assertividade. No entanto, se o número de termos de contexto for muito grande, o algoritmo permite trabalhar com um limiar, através do uso de um outro parâmetro pré-configurado, e assim limitar o número de termos a serem utilizados.

Tendo *score(D ↔ C(Q))* para cada documento recuperado pela máquina de busca, a ordenação de RANKER é obtida a partir do cálculo do ranking decrescente do somatório dos termos descritos no contexto que foram encontrados em cada documento. Por exemplo, a Tabela 1 apresenta a nova ordenação retornada por RANKER juntamente com os respectivos valores calculados para cada um dos documentos, usando como base a consulta apresentada no exemplo da Introdução.

Tabela 1 - Ordenação dos documentos por RANKER com seus respectivos pesos totais.

Num.	Ordenação RANKER	Peso Total
1	http://www.imdb.pt/title/tt0337685/	45.580
2	http://www.museum.tv/eotvsection.php?entrycode=murphybrown	43.537
3	http://www.imdb.pt/title/tt2009259/	30.736
4	http://www.imdb.com/title/tt2009259/	30.736
5	http://www.imdb.com/name/nm0000298/	12.042

É possível notar que, na ordenação proporcionada por RANKER, os documentos retornados pela ordenação da máquina de busca que são, em princípio, mais relevantes para a consulta, aparecem em primeiro lugar, enquanto a ordenação original do motor de busca não garante isso. Comparando com os resultados na Tabela 1, este é o caso, por exemplo, do documento descrito na linha 1 da tabela, que mostrou ser o mais relevante para a consulta e passou para o topo da lista.

Em resumo, a ordenação proporcionada por RANKER através do mecanismo de *blind feedback* é potencialmente mais sensível ao contexto e consequentemente mais relevante para o usuário do que o ranking original da máquina de busca. Essa hipótese será comprovada experimentalmente nos próximos capítulos. A seguir, será apresenta uma nova métrica que permite avaliar de maneira eficiente a relevância posicional de cada documento das listas, tanto as originais fornecidas pelo motor de busca quanto as reordenadas, de forma a comprovar a eficácia de RANKER.

3.5 Métrica *Relative Average Score – RAS*

As seções anteriores iniciaram uma discussão a respeito da relevância dos documentos ordenados por uma máquina de busca como, por exemplo Google, e da necessidade de se investir na concepção de um algoritmo no estilo de RANKER, que seja capaz de retornar para o usuário a lista de documentos selecionados pela máquina em ordem decrescente de relevância. Essa seção esclarece o conceito de relevância

que está sendo considerada no trabalho e formaliza uma nova métrica que se fundamenta nesta noção.

A seção foi dividida em três subseções. A subseção 3.5.1 apresenta a noção de relevância. A subseção 3.5.2 considera o exemplo adotado no capítulo anterior e discute o baixo desempenho da máquina Google considerando a ótica da relevância. A subseção 3.5.3 detalha as tradicionais métricas precisão e revocação. Já a subseção 3.5.4 formaliza a métrica RAS.

3.5.1. Precisão e Revocação

Tradicionalmente, existem duas métricas que permitem avaliar o resultado de uma máquina de busca, conhecidas como precisão e a revocação. Conforme descrito no Capítulo 2, a precisão é a métrica que mede qual o percentual de documentos relevantes no universo de documentos retornados por um motor de busca (MANNING; RAGHAVAN; SCHTZE, 2008). A fórmula da precisão foi apresentada na Equação (2.1). Já a revocação mede qual o percentual de documentos relevantes recuperados por uma pesquisa perante o número total de documentos relevantes na coleção. A fórmula da revocação foi apresentada na Equação (2.2).

A precisão considera todos os documentos retornados pela máquina de busca. Entretanto, uma adaptação bastante utilizada para Web é considerar um ponto de corte nos documentos retornados, devido ao número elevado de resultados retornados. Desse corte, surgem as métricas denominadas P@n (ou TOP-K@n), onde *n* é o ponto de corte.

A revocação é utilizada em conjunto com a precisão para calcular a *Mean Average Precision* (MAP). A métrica MAP unifica em um mesmo escore esse dois valores clássicos da RI (HOFMANN et al., 2009). A métrica é dada pela média das precisões médias (*average precision*) de cada consulta entre as consultas que estão sendo avaliadas. Por sua vez, a precisão média de uma consulta é a média das precisões calculadas após cada documento relevante recuperado. A fórmula que descreve a métrica MAP foi apresentada na expressão (2.3).

Entretanto, o principal problema dessas métricas é que as mesmas não são capazes de distinguir de forma eficiente a relevância posicional, uma vez que operam somente com avaliações binárias [(HOFMANN et al., 2009) (RADLINSKI; CRASWELL, 2010)]. Para contornar esse problema, esse trabalho apresenta também uma nova métrica que suporta avaliação não binária da relevância, a qual foi explicada e formalizada na terceira subseção.

3.5.2. Noção de Grau de Relevância

Como o conceito de significado, relevância é uma noção intuitiva e uma medida determinada não somente pela relação entre documento e pergunta mas também em termos das relações entre os documentos numa relação comparativa (Figueiredo, 1977). A Ciência da Informação tem tentado entender e explicar o fenômeno através de diversas abordagens (matemática, estatística, filosófica) não podendo ser estudada como um fenômeno isolado mas dependente de muitos fatores do processo de comunicação.

Em qualquer lista ordenada (ranking) é importante conseguir avaliar os resultados quantitativamente e de forma eficiente. É preciso calcular a importância relativa dos documentos em cada ranking. Espera-se que os documentos recuperados sejam relevantes, mas em graus variados de relevância dos documentos em relação a consulta realizada ao BD. Conforme indicado no Capítulo 2, o grau de relevância de um documento em relação a uma consulta é função da presença dos resultados da consulta no documento. Mais especificamente, considerou-se que quanto mais intensa a presença do resultado da consulta no documento, maior será o grau de relevância do documento, e que quanto menos intensa, menor o grau.

O baixo desempenho de Google segundo esse aspecto ocorre devido a dois fatores principais. Primeiramente trata-se de um problema de ambiguidade (NAVIGLI, 2009), pois, como pode-se constatar através do próprio Banco de Dados do IMDb, existem 13 títulos de filmes que contem a sub-expressão *‘Murphy Brown’*, portanto Google é incapaz de distinguir *‘Murphy Brown: TV Tales’* dos demais.

Assim, o primeiro, o segundo e o quinto documentos classificados do ranking referem-se respectivamente a outras referências da série Murphy Brown (principalmente de outros produtores).

O segundo fator remete a um problema de reconhecimento de relação entre entidades (MOENS, 2006). O terceiro documento do exemplo do Quadro 1 é uma biografia de *Candice Bergen*, que por sua vez não está diretamente relacionado com a série de TV em questão. Como dito anteriormente, o conceito básico dos algoritmos de ordenação existentes para a recuperação de documentos na WEB é baseado no conceito de relevância geral ou popularidade, ou seja, quanto mais citada é determinada página assim como a qualidade das páginas que cita, maior deverá ser sua ordenação. Assim, a biografia em questão é, de uma maneira geral, extremamente relevante, e a série *Murphy Brown* certamente está citada na mesma, porém não é o foco central do documento.

Em resumo, o Google e outras máquinas de busca, por não ter a ajuda de um contexto de Banco de Dados, que o ajudaria a aliviar os problemas de ambiguidade e reconhecimento de relação entre entidades, não consegue ordenar de forma eficiente determinadas consultas. Esta dissertação visou avaliar a qualidade dos resultados obtidos tanto pela máquina de busca como pela nova lista gerada por RANKER, como forma de evidenciar qual delas é mais aderente ao contexto e consequentemente mais relevante.

3.5.3. Formalização da Métrica RAS

Está implícito na formulação de RAS a noção do grau de relevância de um documento recuperado por uma máquina de busca, em relação a resposta a uma consulta realizada a um BD. Para tal, propõe-se a satisfazer duas condições: (1) quanto mais intensa a presença do resultado da consulta no documento, maior será o grau de relevância do documento, e (2) quanto menos intensa, menor o grau. Esta seção apresenta a formalização de uma métrica de relevância com estas propriedades.

Suponha que doc_i é o i-ésimo documento ordenado em um ranking e PRA_i é a avaliação posicional de relevância (*Positional Relevance Assessment*) desse documento nessa mesma lista. Essa avaliação descreve a posição ideal de um determinado documento em relação ao ranking em questão. A mesma segue o mesmo conceito das demais métricas, com a diferença que aqui o usuário não avaliará somente se o documento é relevante ou não, mas também informará a posição na lista que julga ser a correta para cada documento recuperado.

Mais precisamente, uma avaliação da posição do documento doc_i deve considerar o quão próximo o documento da posição i está de PRA_i. Neste trabalho, a distância entre PRA_i e i é o valor absoluto da diferença, formalmente descrito como $distance_{doc_i} = |i - PRA_i|$. A pontuação dada a um documento em um ranking considera esta distância e o ponto de corte n do ranking, ou seja:

$$score_{doc_i} = (n - distance_{doc_i})/n \tag{3.3}$$

Assim, quando a posição do documento doc_i for considerada ideal, tem-se que $PRA_i = i$, $distance = 0$ e $score = 1$. É importante esclarecer que essa avaliação posicional só se aplica aos documentos avaliados como relevantes. Assim, quando o documento é considerado irrelevante, assume-se PRA = nulo, pois considera-se que a distância é máxima, resultando em $score = 0$.

Considerando a equação (3.3), a métrica *Relative Average Score* (RAS) no ponto de corte n pode ser formalizada da seguinte maneira:

$$RAS@n = \frac{1}{n}\sum_{i=1}^{n} score_{doc_i} \tag{3.3}$$

Quando necessário, médias de *RAS@n* serão indicadas por $\overline{RAS}@n$.

Para efeito de ilustração, vale considerar um comparativo entre Google e RANKER, obtido a partir da avaliação de um especialista quanto aos valores de *PRA* associados aos documentos nos rankings de corte 5 de Google e de RANKER, para a consulta do exemplo apresentado no Capítulo 2. Além dos valores de *PRA*, o Quadro 3 apresenta os valores de *score* individual de cada documento e os valores de *RAS* e *MAP* para Google e RANKER.

Quadro 3 - Comparativo entre a ordenação de Google e a ordenação de RANKER para RAS e MAP.

Num	**Lista ordenada por Google**	***PRA***	***score***	$\overline{RAS}$***@5***	***MAP@5***
1	http://www.imdb.pt/title/tt2009259/	4	0,40	60%	100%
2	http://www.museum.tv/eotvsection.php?entrycode=murphybrown	2	1,00		
3	http://www.imdb.com/name/nm0000298/	5	0,60		
4	http://www.imdb.pt/title/tt0337685/	1	0,40		
5	http://www.imdb.com/title/tt2009259/	3	0,60		
Num	**Lista ordenada por RANKER**	***PRA***	***score***	$\overline{RAS}$***@5***	***MAP@5***
1	http://www.imdb.pt/title/tt0337685/	1	1,00	92%	100%
2	http://www.museum.tv/eotvsection.php?entrycode=murphybrown	2	1,00		
3	http://www.imdb.pt/title/tt2009259/	4	0,80		
4	http://www.imdb.com/title/tt2009259/	3	0,80		
5	http://www.imdb.com/name/nm0000298/	5	1,00		

Conforme a última coluna do Quadro 3 indica, o resultado de *MAP@5* não é capaz de distinguir a relevância relativa entre os documentos, pois considera somente se os mesmos são relevantes ou não. Por outro lado, os valores de *RAS@5* permitem avaliar o quanto a ordenação fornecida por Ranker é melhor do que a fornecida pela máquina de busca. Na presença de várias consultas e resultados para Google e RANKER, $\overline{RAS}$ permite avaliar globalmente cada conjunto de rankings.

4 EXPERIMENTOS

Esse capítulo detalha o plano experimental, que esclarecerá os objetivos dos experimentos, definirá os domínios de teste, ferramentas e avaliadores, as métricas de avaliação e a metodologia utilizada nas avaliações empíricas.

O mesmo está dividido em quatro seções. A 4.1 apresenta o plano experimental das avaliações. Já a seção 4.2 mostra como todas as hipóteses foram validadas e apresenta todos os resultados e gráficos gerados que permitiram analisar os experimentos. A seção 4.3 apresenta o caso do domínio DBLP e suas implicações. Por fim, a seção 4.4 mostra uma comparação entre a nova métrica RAS e MAP.

4.1 Plano Experimental

Os objetivos dos experimentos procuraram validar as seguintes hipóteses:

- H1 – RANKER deve ser competitivo contra um reconhecido trabalho do Estado da Arte;

- H2 – RANKER deve ser bem sucedido em melhorar o ranking dos motores de busca;

- H3 – Embora MAP, uma métrica padrão da RI, remeta a sua maneira para a ordenação dos documentos relevantes, não deve haver nenhuma correlação entre ela e a nova métrica proposta RAS.

Foram escolhidos dois domínios de teste diferentes: filmes e bandas musicais. Em relação ao primeiro domínio, o banco de dados utilizado foi o conhecido *Internet Movie Database* (IMDb). No que diz respeito ao último domínio, a base de dados foi o também conhecido *Open Music Project* - MusicMoz. Em relação às ferramentas utilizadas nos experimentos, SCORE foi o trabalho do Estado

da Arte usado para verificar a hipótese H1 e Google foi o motor de busca utilizado para verificar a hipótese H2.

Além da nova métrica RAS, os padrões em matéria de RI também foram utilizados: MAP e Precisão (TOP-K). Para a hipótese H3, foi escolhido o coeficiente de correlação de Spearman (ou correlação de Spearman, para abreviar). Vale lembrar que a correlação de Spearman entre duas variáveis diminui em magnitude até 0 assim como as duas tornam-se mais distante de ser funções monótonas perfeitas uma da outra.

Especialistas em ambos os domínios foram convidados a sugerir temas livremente, resultado num total de 25 tópicos diferentes por domínio. Da mesma forma, os especialistas escolheram livremente as palavras-chave (termos atômicos do contexto) para as consultas que foram submetidas a Google.

Foram recuperados os 50 primeiros documentos da consulta à Google. Após a execução de RANKER, uma nova lista foi gerada, porém para que o experimento não se tornasse cansativo e os resultados fossem prejudicados, apenas os 10 primeiros documentos de cada lista foram apresentados para a avaliação dos usuários.

A partir dai as duas listas ordenadas foram apresentadas a cada especialista avaliador, em conjunto com os respectivos contextos. Eles foram então convidados a avaliar a qualidade dos documentos, indicando, para cada documento, se o mesmo era relevante para a consulta feita e em que posição o mesmo deveria estar. Em nenhuma circunstância os especialistas foram avisados sobre qual o modelo que resultou na lista de documentos ordenados que estava sendo avaliada naquele momento, se RANKER, Google ou SCORE. Somente lhes foi passado como orientação básica que as avaliações dos documentos deveriam ser feitas com base na adesão ao contexto, mais precisamente na presença dos termos no texto.

Quanto à metodologia, três experimentos E1 - E3 foram planejados da seguinte maneira:

Quadro 4 - Metodologia adotada nos experimentos.

Experimentos E_1 e E_2 (Validação das Hipóteses H_1 e H_2)	
Material	Total de domínios: 2 (IMDb e MusicMoz)
	Total de contextos: 50 (25 para cada domínio)
	Para cada domínio e cada contexto: 1 lista com 10 documentos ordenados por RANKER 1 lista com 10 documentos ordenados por Google
	Total de documentos: 500
Avaliador	5 especialistas, cada um recebendo 10 contextos (5 de cada domínio), num total de 100 documentos para cada um
Métricas	RAS@10, $\overline{\text{RAS}}$@10, $\overline{\text{RAS}}$@2, MAP@10 e TOP-K@1.
Experimento E_3 (Validação da Hipótese H_3)	
Métrica	RAS@10, MAP@10, Correlação de Spearman.

4.2 Experimentos, Gráficos, Tabelas e Análises Correspondentes

É apresentado a seguir, para cada experimento, os respectivos resultados.

4.2.1. Experimento E1

E1 pretende confirmar a hipótese H1, através de uma avaliação experimental comparando RANKER com o algoritmo do Estado da Arte SCORE. Vale lembrar que SCORE se concentra em encontrar as melhores palavras-chave possíveis com o auxílio de um contexto de Banco de Dados, e assumindo que dessa forma o ranking recuperado pela máquina de busca já recupera os documentos na melhor ordem possível.

Abaixo, na Figura 3, temos os resultados alcançados por ambos para RAS@10:

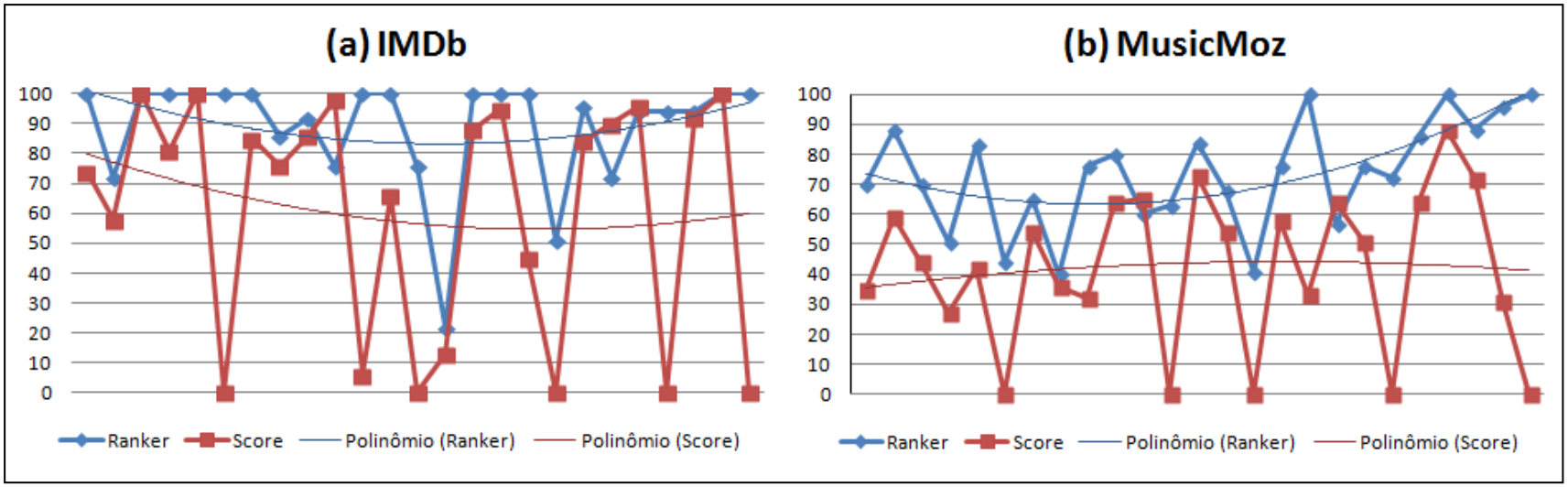

Figura 3 - Resultados de E1 comparando RANKER x SCORE para RAS@10.

Em cada ponto do gráfico de IMDb e MusicMoz, vemos 25 pontos de dados para o valor de RAS@10 e uma linha de tendência polinomial, para RANKER e SCORE respectivamente. É possível observar que os especialistas consideraram vários documentos na lista de SCORE como irrelevantes, deixando o valor da métrica igual a zero. Além disso, os valores de RAS@10 para RANKER são extremamente superiores aos de SCORE, confirmado pela curva de tendência polinomial de cada um. A Tabela 2 abaixo apresenta o resumo dos resultados de E1:

Tabela 2 - Resultado da comparação entre RANKER e SCORE.

Métrica	IMDb		MusicMoz	
	RANKER	**SCORE**	**RANKER**	**SCORE**
$\overline{RAS}$@10	93,4%	69,3%	73,4%	41,8%
$\overline{RAS}$@2	90,0%	55,0%	88,0%	34,0%
MAP@10	90,3%	75,1%	75,0%	38,7%
TOP-K@1	84,0%	48,0%	96,0%	32,0%

Os valores de $\overline{RAS}$@10 e $\overline{RAS}$@2 mostram, em IMDb, uma vantagem pró RANKER de 24,1% e 35% respectivamente, e em MusicMoz, uma vantagem de 31,6% e incríveis 54% respectivamente. O valor de *k* igual a 2 para RAS, assim como a métrica TOP-K@1 são excelentes indicativos para identificar se os primeiros documentos recuperados são realmente relevantes e estão de acordo com suas relevâncias posicionais. E em ambos se percebe que RANKER cumpriu seu papel, melhorando a ordenação da lista e trazendo os documentos mais relevantes para o topo.

4.2.2. Experimento E2

O experimento E2 coloca RANKER frente a frente com a máquina de busca mais utilizada da atualidade, Google[7]. Diferentemente de E1, aqui o conjunto de palavras-chave enviado foi o mesmo, ou seja, aquelas geradas por RANKER. A Figura 4 apresenta um gráfico comparativo do resultado conseguidos pelos rankings de RANKER e Google tanto para IMDb quanto para MusicMoz, usando a métrica RAS@10:

[7] Pesquisa disponível em: http://searchenginewatch.com/article/2067079/Nielsen-NetRatings-Search-Engine-Ratings

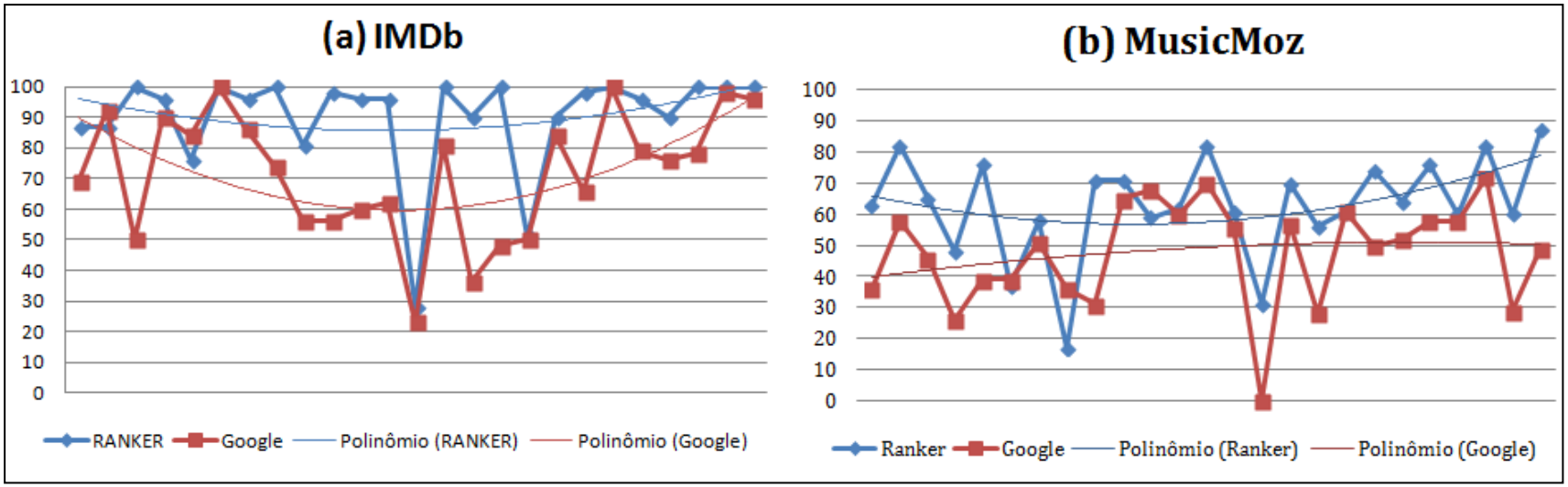

Figura 4 - Resultados de E2 comparando RANKER x Google para RAS@10.

Da mesma forma, cada gráfico apresenta 25 pontos com os resultados de RAS@10 alcançados por RANKER e Google, além de uma linha de tendência polinomial. O que observamos é que os valores de RAS@10 para RANKER são constantemente superiores as de Google, mesmo que em algumas vezes com pequena margem de diferença. A Tabela 3, a seguir, resumo os resultados do experimento e apresenta os valores para as demais métricas consideradas:

Tabela 3 - Resultado da comparação entre Ranker e Google.

Métrica	IMDb		MusicMoz	
	Ranker	**Google**	**Ranker**	**Google**
$\overline{\text{RAS}}$@10	90,2%	72,8%	62,7%	47,8%
$\overline{\text{RAS}}$@2	90,0%	60,0%	85,0%	68,0%
MAP@10	90,3%	75,1%	66,4%	54,6%
TOP-K@1	88,0%	52,0%	76,0%	28,0%

Assim como em E1, os resultados a favor de RANKER foram extremamente satisfatórios, com margem em IMDb de 17,4% e 30% para $\overline{\text{RAS}}$@10 e $\overline{\text{RAS}}$@2 respectivamente, e para MusicMoz 14,9% e 17% respectivamente. Da mesma forma, $\overline{\text{RAS}}$@2 e TOP-K@1 evidenciam o poder de RANKER em trazer os documentos mais relevantes para as primeiras posições das listas.

Os resultados apresentados mostram que, tanto para o domínio IMDb quanto para MusicMoz, a diferença dos valores de $\overline{\text{RAS}}$ a favor de RANKER é estatisticamente significativa, sempre maiores que 5% (JAIN, 1991). Além disso, a métrica de precisão TOP-K@1 retrata que a vantagem de RANKER para ambos os domínios também é clara: 36 pontos percentuais para IMDb e 48 pontos percentuais para MusicMoz. Considerando estes resultados, a conclusão é que as hipóteses E1 e E2 estão validadas experimentalmente.

4.2.3. Experimento E3

Esse experimento destaca os valores da correlação de Spearman, comparando as métricas RAS e MAP através da pontuação de RANKER e Google, em relação aos domínios IMDb e MusicMoz. Os mesmos estão expostos no Quadro 5, e apresentam os valores para o cálculo das correlações de RAS@10 e MAP@10:

Quadro 5 - Correlação de Spearman entre RAS e MAP.

Domínio	Comparação	Correlação de Spearman
IMDb	RANKER vs. Google	0.54768
IMDb	RANKER vs. SCORE	0.59763
MusicMoz	RANKER vs. Google	0.71700
MusicMoz	RANKER vs. SCORE	0.66884

Todos os valores de correlação do Quadro 5 são baixos ou muito abaixo do limite superior de 8,0 normalmente utilizado, o que significa que não há nenhuma correlação entre RAS e MAP. Assim, considerando estes resultados, a conclusão é que a hipótese H3 está validada.

4.3 O Caso DBLP

Além dos dois domínios descritos nos experimentos, um outro domínio também foi estudado. Trata-se do *Digital Bibliography & Library Project* (DBLP)[8], uma conceituada base que relaciona os principais pesquisadores da área de BD com suas respectivas publicações. O mesmo foi submetido para a avaliação de especialistas da mesma forma que os anteriores, porém os resultados apresentados forma divergentes e não conclusivos.

Para a hipótese H1, os resultados continuaram mostrando clara vantagem de RANKER frente SCORE, como demonstrado na Figura 5:

[8] Script available on: http://kdl.cs.umass.edu/data/dblp/dblp-info.html

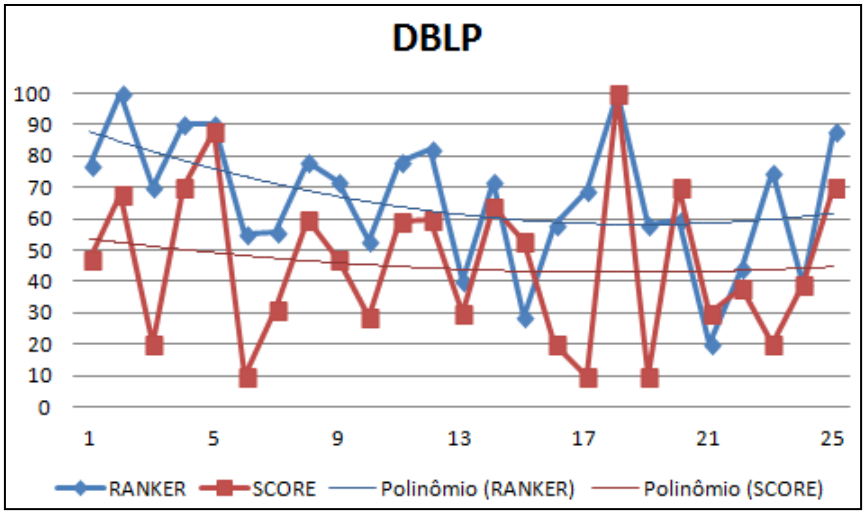

Figura 5 - Resultados de DBLP para RANKER x SCORE.

A figura mostra que RANKER foi constantemente mais bem avaliado que SCORE, perdendo em apenas 3 (três) temas e empatando em outros 2 (dois). Esse fato também foi confirmado pelas curvas polinomiais de cada conjunto.

Entretanto, os resultados do domínio DBLP para RANKER x Google não foram tão conclusivos quanto os de SCORE. A Figura 6 mostra um gráfico que apresenta os respectivos desempenhos:

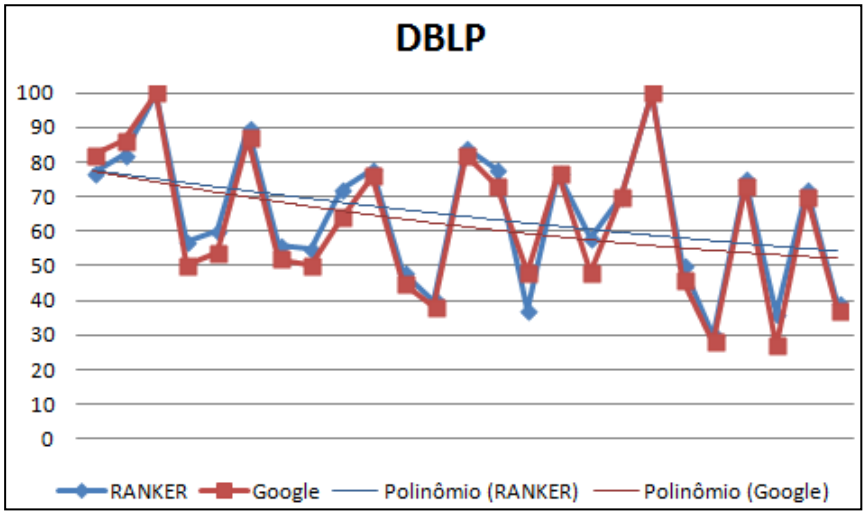

Figura 6 - Resultados de DBLP para RANKER x Google.

A partir dos resultados não é possível concluir que RANKER se saiu melhor, visto que a diferencia média ficou em aproximadamente 2%.

Fazendo uma análise mais detalhada, é possível entender os motivos que levaram a avaliação a esses resultados. Usaremos como exemplo o seguinte tema: *"Quais as publicações de Hugo Zaragoza?"*. Por se tratar de um autor de peso da área, o mesmo possui muitas publicações, portanto o resultado dessa consulta ao BD retorna muitas linhas (51 para a base utilizada), dificultando a escolha das palavras-chave pelos usuários. Esse fato levaram os conjuntos de termos submetidos às máquinas de busca a ficarem muito genéricos, prejudicando a qualidade dos resultados.

Além disso, dos 10 documentos mais bem ranqueados, somente 3 referenciaram o literal *"Hugo Zaragoza"*, sendo que os 7 demais não tinham qualquer ligação com o autor. Esse fato levou ao mal desempenho tanto de RANKER quanto de Google, e a reordenação não teve efeito sobre o resultado final. Vale ressaltar que vários documentos não referenciam o literal completo, mas sim por citações que variam bastante, como no caso em questão temos "Zaragoza, H.," "Zaragoza, Hugo", "Zaragoza et al.", entre outros. Tal fato também contribuiu para prejudicar a reordenação.

DBLP possui uma característica que ajuda a explicar esses problemas. Na mesma dificilmente o usuário conseguirá escolher bem os termos do contexto para compor o conjunto de palavras-chave e essa escolha, pois qualquer que seja, deixará a busca vaga e com alta probabilidade de retornar documentos irrelevantes. A grande diferença para os demais domínios é que nos demais casos os termos do contexto e as palavras-chave da maquina de busca são praticamente os mesmos, facilitando o casamento e consequente encontro de documentos relevantes.

Portanto, DBLP serviu para identificarmos uma limitação do algoritmo, que é lidar com contextos muito grandes e com pouquíssimos termos com altos pesos. Nesse cenário, os resultados das máquinas de busca dificilmente poderão ser reordenados de forma eficiente, fazendo com que as listas se mantenham praticamente iguais.

4.4 Discussão Comparativa Entre RAS e MAP

A Tabela 4, apresentada abaixo, a qual foi extraída dos resultados gerais dos experimentos (apresentados em detalhes nos anexos de A a D), deixa claro a deficiência de MAP e os motivos que levaram a necessidade da criação da nova métrica RAS, que viabiliza a avaliação da relevância posicional de cada documento.

Tabela 4 - Comparativo resultados RAS e MAP no IMDb para RANKER e Google.

Filme IMDb	RAS@10		MAP@10	
	RANKER	Google	RANKER	Google
Alexis Zorbas	100%	50%	100%	100%
Apocalypse Now	96%	90%	100%	100%
Good Will Hunting	96%	86%	100%	100%
Teen Choice Awards 2003, The	98%	66%	100%	100%
Women in Love	100%	96%	100%	100%

Os temas em questão tiveram todos os documentos avaliados como relevantes pelos especialistas, porém a posição em que os mesmos se encontravam nas listas não estavam corretas. Assim, MAP, por não ser capaz de distinguir a relevância posicional dos documentos, não permite avaliar com qualidade os resultados. RAS, por sua vez é capaz de capturar essa característica e nos casos citados quantifica de forma precisa a melhora proporcionada por RANKER.

5 CONCLUSÃO E TRABALHOS FUTUROS

A principal contribuição do trabalho foi a proposta de um novo método para efetuar a integração BD-RI, batizado de RANKER, que funciona da seguinte maneira: Um usuário que faz uma consulta a uma base de dados tem ainda a opção de receber informações relevantes extras a partir de fontes de textos não estruturados e que são apresentadas em ordem aproximada de relevância para a consulta. O método se baseia em três etapas:

- O usuário submete uma consulta a um SGBD;

- O usuário é convidado a opcionalmente indicar palavras-chave que julgue relevantes do conjunto-resposta da consulta que, juntamente com os literais dessa mesma consulta, são submetidos para um motor de busca;

- Sem que o usuário saiba, RANKER reordena os documentos devolvidos pelo motor de busca de acordo com um contexto da consulta à base de dados, sob a forma de um mecanismo de *blind feedback*. A preocupação é, por conseguinte, a eficácia da solução de integração BD-RI.

Como as métricas tradicionais da RI não são capazes de distinguir bem a relevância posicional dos documentos de uma coleção, uma vez que trabalham com operações binárias (relevante ou não), uma nova métrica foi criada e que leva em conta a avaliação de relevância posicional dos documentos em uma lista. A mesma foi batizada de RAS (*Relative Average Score*) e permite quantificar a qualidade da ordenação fornecida por uma máquina de busca.

A viabilidade do modelo de integração proposto foi evidenciada por meio de avaliações empíricas. Até onde avançaram nossas pesquisas, não foi encontrada nenhuma outra solução de integração BD-RI que tenha atingido o nível de qualidade aqui apresentado.

Como trabalho futuro, pretende-se ainda aprimorar a parte de seleção dos termos a serem submetidos para a máquina de busca, através de uma pré-seleção automatizada dos termos mais relevantes de acordo com a consulta e a relevância para

a coleção em si, para somente a partir desse momento apresentar ao usuário que efetuará a escolha. Essa melhoria poderá ajudar a resolver os problemas encontrados para o domínio do DBLP.

Um outro tópico interessante a ser abordado como trabalho futuro é a questão de como estender o modelo para apoiar RANKER em consultas que não sejam conjuntivas. Por fim também pretende-se melhorar a eficácia do modelo para reordenar os documentos, melhorando a sua capacidade para o reconhecimento das relações entre as entidades assim como reconhecimento de ambiguidade com a ajuda do contexto das consultas às base de dados.

BIBLIOGRÁFICA CONSULTADA

Abiteboul S. et al. The Lowell Database: Research Self-Assessment. Communications of the ACM, Vol. 48, No. 5, 2005.

Agrawal R. et al. The Claremont Report on Database Research. SIGMOD Record, September 2008 (Vol. 37, No. 3).

Amer-Yahia S., Hiemstra D., Roelleke T., Srivastava D., Weikum G. DB&IR Integration: Report on the Dagstuhl Seminar "Ranked XML Querying". SIGMOD Record, September 2008 (Vol. 37, No. 3).

Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley/ACM, 1999.

Chakaravarthy V. T., Gupta H., Roy P., Mohania M. Efficiently Linking Text Documents with Relevant Structured Information. *VLDB '06,* September 12-15, 2006.

Chen L. J., Papakonstantinou Y. Context-sensitive Ranking for Document Retrieval. SIGMOD'11, June12–16, 20.

Doan A., Halevy A. Y. Semantic-integration research in the database community. AI Mag. 26, 1 (March 2005), 83-94, 2005.Figueiredo, L. M. de. O conceito de relevância e suas implicações. Ciência da Informação, Rio de Janeiro, v. 6, n. 2. p. 75-78, 1977.

Fagin R., Kimelfeld B., Li Y., Raghavan S., Vaithyanathan S. Understanding Queries in a Search Database System. PODS'10, June 6–11, 2010.

Fagin R., Kimelfeld B., Li Y., Raghavan S., Vaithyanathan S. Rewrite Rules for Search Database Systems. PODS'11, June 13–15, 2011.

Gruber, T. R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2):199-220, 1993.

Hristidis V., Gravano L., Papakonstantinou Y. Efficient IR-style keyword search over relational databases. In Proceedings of the 29th international conference on Very large data bases - Volume 29 (VLDB '03). VLDB Endowment 850-861, 2003.

Jain, R. K. The Art of Computer Systems Performance Analysis: Techniques for Design, Measurement, Simulation, and Modeling. Wiley, 1991.

Järvelin K., Kekäläinen J. Cumulated Gain-based Evaluation of IR Techniques. ACM TOIS, Vol. 20, No. 4, October 2002, pp. 422-446.

Katja Hofmann, Manos Tsagkias, Edgar Meij, and Maarten de Rijke. The impact of document structure on keyphrase extraction. In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09). ACM, New York, NY, USA, 1725-1728, 2009.

Manning C. D., Raghavan P., Schtze P. *Introduction to Information Retrieval*. P. 193-194. Cambridge University Press, New York, NY, USA, 2008.

Moens M-F. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, 2006.

Mukesh K. Mohania, Manish Bhide: New trends in information integration. ICUIMC, pp. 74-81, 2008.

Navigli R. Word Sense Disambiguation: a Survey. ACM Computing Surveys, Vol. 41, No. 2, 2009.

Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report n. 1999-66, Stanford University, Stanford, 1999.

Radlinski F., Craswell N. Comparing the Sensitivity of Information Retrieval Metrics, SIGIR’10, July 19–23, 2010.

Roy P., Mohania M., Bamba B., Raman S. Towards Automatic Association of Relevant Unstructured Content with Structured Query Results. CIKM’05, October 31–November 5, 2005.

Whang K-Y. DB-IR Integration and Its Application to Massively-Parallel Search Engine, CIKM´09, 2009.

Whang K-Y. et al. Odysseus: A High-Performance ORDBMS Tightly-Coupled with IR Features, 21st International Conference on Data Engineering (ICDE'05), pp. 1104-1105, 2005.

ANEXOS

ANEXO A – RESULTADOS DE IMDb PARA RANKER x SCORE

Num	Filme	Avaliador	Algoritmo					
			RAS@10		RAS@2		MAP@10	
			RANKER	SCORE	RANKER	SCORE	RANKER	SCORE
1	10 Items or Less	Avaliador 01	100	74	100	50	100	80
2	50 Greatest Comedy Films, The	Avaliador 01	72	58	100	50	78	45
3	Alexis Zorbas	Avaliador 01	100	100	100	100	100	100
4	Apocalypse Now	Avaliador 01	100	81	100	50	100	88
5	Cotton Club, The	Avaliador 01	100	100	100	100	100	100
6	Godfather: Part III, The	Avaliador 02	100	0	100	0	100	0
7	Good Will Hunting	Avaliador 02	100	85	100	50	100	90
8	Intolerable Cruelty	Avaliador 02	86	76	100	100	100	100
9	King of Comedy, The	Avaliador 02	92	86	100	100	100	100
10	Ladykillers, The	Avaliador 02	76	98	100	100	100	100
11	Mexican, The	Avaliador 03	100	6	100	0	100	62
12	Minority Report	Avaliador 03	100	66	50	25	90	100
13	Murphy Brown: TV Tales	Avaliador 03	76	0	100	0	100	0
14	Queimada	Avaliador 03	22	13	0	0	5	2
15	Raging Bull	Avaliador 03	100	88	100	0	100	100
16	Scarface	Avaliador 04	100	95	100	100	100	90
17	Star Wars: Episode III	Avaliador 04	100	45	100	50	100	51
18	Sunday Bloody Sunday	Avaliador 04	51	0	50	0	48	0
19	Teen Choice Awards 2003, The	Avaliador 04	96	84	100	100	100	100
20	Terminator 2: Judgment Day	Avaliador 04	72	90	50	100	100	100
21	That Thing You Do!	Avaliador 05	94	96	100	100	100	80
22	The Birds	Avaliador 05	94	0	100	0	100	0
23	Ultimo Tango a Parigi	Avaliador 05	94	92	100	100	100	100
24	Viva Zapata!	Avaliador 05	100	100	100	100	100	100
25	Women in Love	Avaliador 05	100	0	100	0	100	0
Total			**89,0**	**61,3**	**90,0**	**55,0**	**92,8**	**67,5**

ANEXO B – RESULTADOS DE MUSICMOZ PARA RANKER x SCORE

Num	Filme	Avaliador	Algoritmo					
			RAS@10		RAS@2		MAP@10	
			RANKER	SCORE	RANKER	SCORE	RANKER	SCORE
1	Aerosmith	Avaliador 01	70	35	100	0	73	24
2	Bee Gees	Avaliador 01	88	59	100	50	90	52
3	Bob Marley	Avaliador 01	70	44	50	50	80	38
4	Black Sabbath	Avaliador 01	51	27	50	50	40	23
5	Creedence Clearwater Revival	Avaliador 01	83	42	100	50	84	33
6	Dave Matthews Band	Avaliador 02	44	0	50	0	38	0
7	INXS	Avaliador 02	65	54	100	50	67	52
8	Led Zeppelin	Avaliador 02	40	36	50	50	35	40
9	Linking Park	Avaliador 02	76	32	100	0	76	21
10	Metallica	Avaliador 02	80	64	100	0	80	51
11	Nirvana	Avaliador 03	60	65	100	25	55	56
12	Oasis	Avaliador 03	63	0	50	0	73	0
13	Offspring	Avaliador 03	84	73	100	100	82	69
14	Pearl Jam	Avaliador 03	68	54	100	50	78	62
15	Phoenix	Avaliador 03	41	0	50	0	40	0
16	Pink Floyd	Avaliador 04	76	58	100	100	80	90
17	Queen	Avaliador 04	100	33	100	50	100	25
18	Question Mark and the Mysterians	Avaliador 04	57	64	100	0	79	51
19	Red Hot Chili Peppers	Avaliador 04	76	51	100	100	75	59
20	Rush	Avaliador 04	72	0	100	0	75	0
21	The Beach Boys	Avaliador 05	86	64	100	0	85	51
22	The Beatles	Avaliador 05	100	88	100	100	100	88
23	The Eagles	Avaliador 05	88	72	100	25	90	62
24	The Rolling Stones	Avaliador 05	96	31	100	0	100	21
25	Wings	Avaliador 05	100	0	100	0	100	0
Total			**73,4**	**41,8**	**88,0**	**34,0**	**75,0**	**38,7**

ANEXO C – RESULTADOS DE IMDb PARA RANKER x GOOGLE

Num	Filme	Avaliador	Algoritmo					
			RAS@10		RAS@2		MAP@10	
			Ranker	Google	Ranker	Google	Ranker	Google
1	10 Items or Less	Avaliador 01	87	69	100	25	87	59
2	50 Greatest Comedy Films, The	Avaliador 01	87	92	100	100	87	82
3	Alexis Zorbas	Avaliador 01	100	50	100	50	100	100
4	Apocalypse Now	Avaliador 01	96	90	100	100	100	100
5	Cotton Club, The	Avaliador 01	76	84	100	100	76	85
6	Godfather: Part III, The	Avaliador 02	100	100	100	100	100	100
7	Good Will Hunting	Avaliador 02	96	86	100	100	100	100
8	Intolerable Cruelty	Avaliador 02	100	74	100	100	100	72
9	King of Comedy, The	Avaliador 02	81	56	25	0	71	51
10	Ladykillers, The	Avaliador 02	98	56	100	50	100	50
11	Mexican, The	Avaliador 03	96	60	100	0	100	51
12	Minority Report	Avaliador 03	96	62	100	0	100	51
13	Murphy Brown: TV Tales	Avaliador 03	28	23	0	50	18	20
14	Queimada	Avaliador 03	100	81	100	25	100	71
15	Raging Bull	Avaliador 03	90	36	100	0	90	26
16	Scarface	Avaliador 04	100	48	100	50	100	82
17	Star Wars: Episode III	Avaliador 04	51	50	25	0	48	48
18	Sunday Bloody Sunday	Avaliador 04	90	84	100	0	90	85
19	Teen Choice Awards 2003, The	Avaliador 04	98	66	100	50	100	100
20	Terminator 2: Judgment Day	Avaliador 04	100	100	100	100	100	100
21	That Thing You Do!	Avaliador 05	96	79	100	100	100	87
22	The Birds	Avaliador 05	90	76	100	100	90	78
23	Ultimo Tango a Parigi	Avaliador 05	100	78	100	100	100	78
24	Viva Zapata!	Avaliador 05	100	98	100	100	100	100
25	Women in Love	Avaliador 05	100	96	100	100	100	100
Total			**90,2**	**71,8**	**90,0**	**60,0**	**90,3**	**75,1**

ANEXO D – RESULTADOS DE MUSICMOZ PARA RANKER x GOOGLE

MusicMoz								
Num	**Banda**	**Avaliador**	**Algoritmo**					
			RAS@10		**RAS@2**		**MAP@10**	
			Ranker	**Google**	**Ranker**	**Google**	**Ranker**	**Google**
1	Aerosmith	Avaliador 01	63	36	100	50	63	29
2	Bee Gees	Avaliador 01	82	58	100	50	100	62
3	Bob Marley	Avaliador 01	65	46	100	50	84	57
4	Black Sabbath	Avaliador 01	48	26	50	25	31	27
5	Creedence Clearwater Revival	Avaliador 01	76	39	100	50	66	41
6	Dave Matthews Band	Avaliador 02	37	39	25	100	37	36
7	INXS	Avaliador 02	58	51	100	100	50	57
8	Led Zeppelin	Avaliador 02	17	36	50	50	10	31
9	Linking Park	Avaliador 02	71	31	100	50	67	28
10	Metallica	Avaliador 02	71	65	100	100	73	62
11	Nirvana	Avaliador 03	59	68	100	100	50	89
12	Oasis	Avaliador 03	62	60	50	100	60	58
13	Offspring	Avaliador 03	82	70	100	100	100	79
14	Pearl Jam	Avaliador 03	61	56	100	50	47	51
15	Phoenix	Avaliador 03	31	0	50	0	39	0
16	Pink Floyd	Avaliador 04	70	57	100	100	72	78
17	Queen	Avaliador 04	56	28	50	50	58	34
18	Question Mark and the Mysterians	Avaliador 04	61	61	50	100	84	60
19	Red Hot Chili Peppers	Avaliador 04	74	50	100	100	89	60
20	Rush	Avaliador 04	64	52	100	50	73	73
21	The Beach Boys	Avaliador 05	76	58	100	25	100	74
22	The Beatles	Avaliador 05	60	58	100	100	100	74
23	The Eagles	Avaliador 05	82	72	100	100	82	100
24	The Rolling Stones	Avaliador 05	60	29	100	0	44	32
25	Wings	Avaliador 05	87	49	100	100	84	75
Total			**62,9**	**47,8**	**85,0**	**68,0**	**66,4**	**54,6**