Recuperação de Informação: Visão Geral

Rodiney Elias Marçal[1], **Liliane do Nascimento Vale**[2]

[1]Universidade Federal de Goiás - CAC
Departamento de Ciência da Computação
Catalão/GO, Brasil

[2]Universidade Federal de Goiás - CAC
Departamento de Ciência da Computação
Catalão/GO, Brasil

rodiney@gmail.com, liliane.ufg@gmail.com

Abstract. *Over time, man has been accumulating large amounts of documents capable of transmitting information. Retrieve information simply means finding a set of documents that is relevant to a user's needs. The area of information retrieval is a relatively new area in computer science that deals with the representation, storage, organization and access to objects (documents) information. The purpose of this paper is to approach, in a more general way, the field of information retrieval along with its definition, architecture, techniques and applications.*

Resumo. *Ao longo do tempo, o ser humano vem acumulando grandes quantidades de documentos capazes de transmitir informação. Recuperar informação simplesmente significa encontrar um conjunto de documentos que seja relevante a uma necessidade do usuário. A área de* Recuperação de Informação *(*Information Retrieval*) é uma área relativamente nova na ciência da computação que lida com a representação, armazenamento, organização e acesso a objetos (documentos) de informação. O intuito desse artigo é explorar de maneira geral o campo da Recuperação de Informação, apresentando sua definição, arquitetura, técnicas e aplicações.*

1. Introdução

Armazenar e recuperar informações é uma necessidade antiga do homem e, com os avanços da tecnologia, está cada vez maior a quantidade de informações disponíveis. Essa disponibilidade influenciou para o surgimento da área de Recuperação de Informação (RI).

Recuperação de Informação (RI) é a área que lida com a representação, busca e manipulação de grandes coleções de texto eletrônico e outros dados relacionados com a linguagem humana [Büttcher et al. 2010].

Sistemas de recuperação de informação, ou simplesmente sistemas de RI, possibilitam a seus usuários o acesso a grandes quantidades de dados armazenados eletronicamente. Assim, o usuário que submete uma consulta em um sistema de recuperação de informação receberá como resposta uma série de documentos relacionados com a sua solicitação [Ruthven and Lalmas 2003]. Vale mencionar que o termo *documento* possui

um significado mais amplo, se referindo a qualquer unidade que pode ser retornada para o usuário como resultado de uma busca. Na prática, então, um documento pode ser uma mensagem de e-mail, uma página da Web, uma imagem, ou mesmo um vídeo.

Recuperação de informação é a base para os motores de busca modernos. Esse artigo analisa os conceitos envolvidos para a implementação de um motor de busca.

2. Recuperação de Informação

Recuperar informações simplesmente significa encontrar um conjunto de documentos que seja relevante a uma consulta do usuário.

Algumas características diferenciam um sistema de Recuperação de Informação (RI) de uma ferramenta de acesso a informações. Por exemplo, um sistema de RI não extrai informação dos objetos obtidos por ele. Além disso, sistemas de RI não realizam nenhum tipo de processamento de informação contida dentro desses objetos. São esses pontos que separam um sistema de recuperação de informações de sistemas baseados em conhecimento, tais como os sistemas especialistas, grafos conceituais ou redes semânticas. Esses sistemas baseados em conhecimentos dependem massivamente da representação pré-definida de um domínio, tal como a medicina ou a advocacia. Dessa forma, o conhecimento desse domínio pode ser usado para manipular, inferir ou categorizar informações para um usuário. Por outro lado, sistemas de RI são usados para direcionar o usuário aos objetos que possam ajudá-lo a satisfazer a sua necessidade de informação [Ruthven and Lalmas 2003].

Serviços que empregam a recuperação de informação estão ficando cada vez mais difundidos, com milhões de usuários dependendo deles diariamente para tratar de negócios, educação e entretenimento. Mecanismos de buscas na Web - Google, Bing, e outros - são de longe os serviços de RI mais populares e utilizados.

3. Arquitetura de um sistema de RI

A maioria dos sistemas de RI compartilham de uma mesma organização e arquitetura básicas que são adaptadas conforme os requisitos específicos de cada aplicação. A Figura 1 ilustra os principais componentes de um sistema de RI.

Antes de realizar uma busca, um usuário possui uma *necessidade de informação*, a qual sustenta e motiva o processo de pesquisa. Essa necessidade de informação muitas vezes é referenciada como sendo um *tópico*, especialmente quando ela é apresentada de forma escrita como parte de um conjunto de testes para avaliação de um sistema de RI. Como resultado dessa necessidade de informação, o usuário constrói e emite uma consulta (*query*) ao sistema de recuperação de informação. Tipicamente, essa consulta consiste de um pequeno número de *termos*, com dois ou três termos principalmente quando se trata de uma busca voltada para a Web. A designação *termo* é utilizada ao invés de *palavra* porque um termo de uma consulta não necessariamente precisa significar uma palavra. Assim, dependendo da necessidade de informação, um termo da consulta poder ser uma data, um número, uma nota musical ou mesmo uma imagem. Operadores curinga também podem ser permitidos como termos da consulta. Por exemplo, o termo *inform** pode casar com qualquer palavra iniciando-se com esse prefixo, isto é, *inform*, *informação*, *informal*, *informante*, *informativo*, etc [Büttcher et al. 2010].

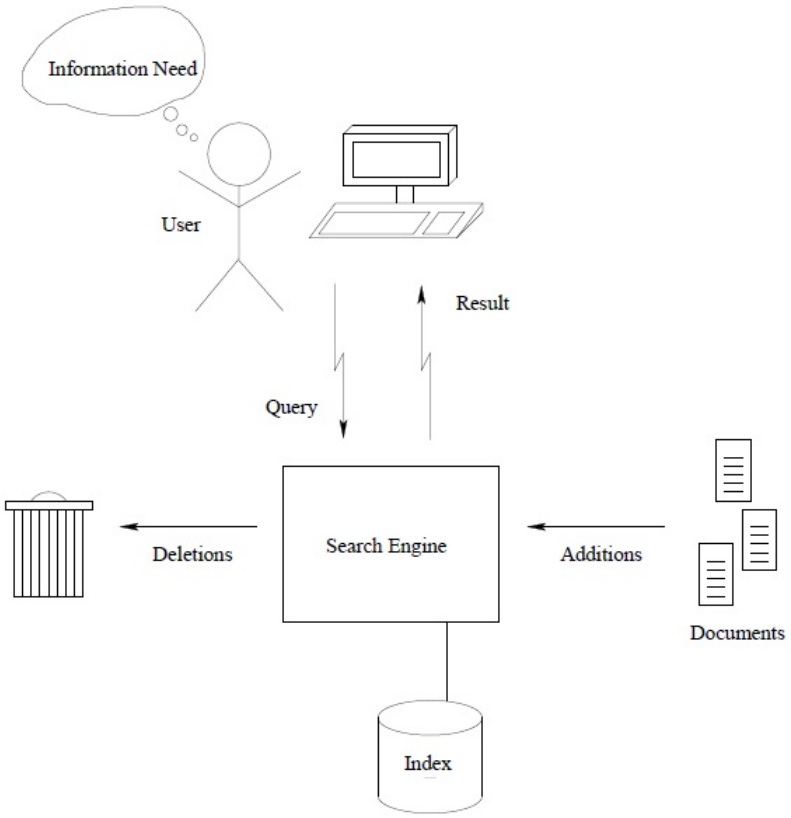

Figura 1. Componentes de um sistema de RI [Büttcher et al. 2010]

Embora usuários comumente empregam palavras-chaves simples para a formulação das consultas, a maioria dos sistemas de RI possuem suporte a uma sintaxe mais rica e ampla, tal como expressões booleanas e operadores de casamento de padrão. Essas facilidades podem ser utilizadas, por exemplo, para limitar a busca apenas para um determinado website ou um domínio específico [Büttcher et al. 2010].

A consulta do usuário é então processada por um mecanismo de busca (*search engine*), o qual pode estar sendo executado em uma máquina local, ou em um ambiente de clusters em uma localização geográfica remota, ou qualquer outro lugar.

Uma das principais tarefas de um mecanismo de busca é manter e manipular um *índice invertido* para uma coleção de documentos. Esse índice forma a principal estrutura de dados utilizada pelo mecanismo para a realização da busca e classificação por relevância. A função básica de um índice invertido é fornecer um mapeamento entre termos e a sua localização nos documentos em que eles aparecem [Büttcher et al. 2010].

Para suportar algoritmos de classificação por relevância, o mecanismo de busca mantém estatísticas associadas com o índice, tais como o número de documentos contendo cada termo e o comprimento de cada documento. Além disso, o mecanismo de busca geralmente possui acesso ao conteúdo original do documento, a fim de informar resultados significativos de volta para o usuário[Büttcher et al. 2010].

De posse do índice invertido, da coleta de estatísticas e outros dados, o mecanismo de busca aceita as consultas do usuário, as processa, e então retorna uma lista de resulta-

dos classificados. Para executar uma classificação por relevância, o mecanismo de busca calcula uma pontuação (*score*) para cada documento. Após ordenar os documentos de acordo com suas pontuações, a lista de resultados pode ser tratada, onde registros duplicados ou redundantes podem ser removidos. Por exemplo, um mecanismo de busca na Web poderia apresentar apenas um ou dois resultados de um único domínio, eliminando os outros para favorecer páginas de fontes distintas [Büttcher et al. 2010]. O problema de computar o *score* de um documento com relação a consulta do usuário constitui um dos problemas mais importantes no campo da RI.

4. Indexação

Um sistema de RI desempenha sua tarefa através da indexação dos documentos (a menos que o sistema utilize o documento diretamente) e da reformulação de consultas, resultando assim, respectivamente, na representação de documentos e representação das consultas. O sistema então efetua os casamentos (*matches*) sobre a representação e exibe os documentos que são encontrados, permitindo ao usuário a seleção dos itens relevantes. Além disso, a busca pode passar por várias iterações, sendo que a análise a respeito das características que distinguem documentos relevantes de não-relevantes pode ser usada para melhorar a consulta ou a indexação, tal como proposto pelo *feedback de relevância* [Ruthven and Lalmas 2003].

Para pequenos conjuntos de documentos é possível que um sistema de RI avalie um documento de cada vez, decidindo se o documento em avaliação é ou não relevante de acordo com a consulta do usuário. Entretanto, para conjuntos com grande quantidade de documentos, especialmente no caso de sistemas interativos, esta abordagem torna-se impraticável. Por isso, faz-se necessário modificar o conjunto original de documentos para uma representação facilmente acessível: uma que seja capaz de referenciar os documentos mais prováveis de serem relevantes, por exemplo aqueles que contenham pelo menos uma palavra que apareça na consulta do usuário [Ruthven and Lalmas 2003].

A transformação de um texto de um documento em uma representação textual é conhecida como indexação. Há uma variedade de técnicas de indexação do documento, mas a maioria delas se baseiam na seleção de bons descritores para os documentos, tais como palavras-chaves (ou *termos*) usados para representar o conteúdo da informação contida nos documentos. Um ”bom”descritor para a RI é um termo que ajuda a descrever o conteúdo das informações no documento e que também ajuda a diferenciar um documento dos outros documentos do conjunto. Um ”bom”descritor, portanto, possui um certo nível discriminatório. Esse nível pode ser usado na diferenciação entre documentos relevantes e não-relevantes [Ruthven and Lalmas 2003].

A Figura 2 esboça os passos básicos na transformação de um documento para sua forma indexada. O primeiro estágio é converter o texto do documento (Texto do documento - Figura 2 - a) em um fluxo (*stream*) de termos, normalmente convertendo todos os termos para letras minúsculas e removendo os caracteres de pontuação (*Tokenização* - Figura 2 - b).

Uma vez que o texto do documento tenha sido tokenizado, é necessário decidir quais termos deverão ser usados para representar o documento. Portanto, faz-se necessário decidir quais descritores serão úteis tanto para descrever o conteúdo do documento quanto para discriminar esse documento de outros documentos do conjunto.

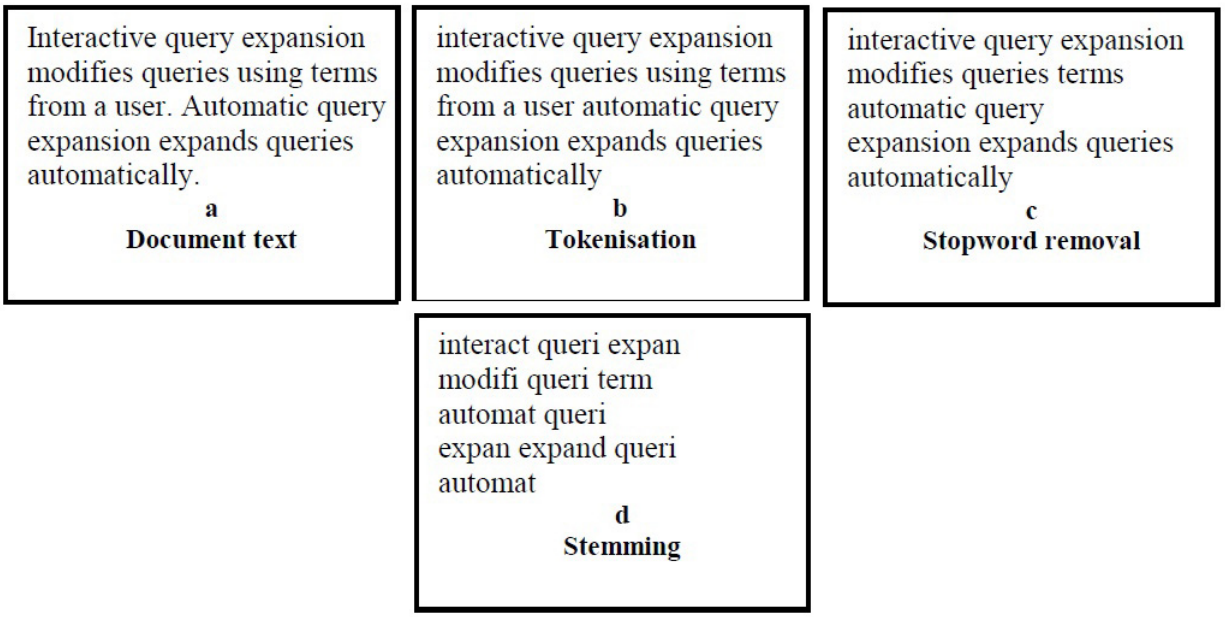

Figura 2. Indexação de um documento [Ruthven and Lalmas 2003]

Termos de elevada frequência, ou seja, aqueles que aparecem em grande proporção no documento, tendem a não ser efetivos para a discriminação e nem para a representação [Ruthven and Lalmas 2003].

Existem duas razões principais para isso. A primeira é que, para a maioria das consultas reais dos usuários, é provável que o número de documentos relevantes para uma consulta representa uma pequena proporção dos documentos do conjunto. Assim, também é provável que um termo que consiga separar documentos relevantes de documentos não-relevantes será um termo que aparece somente em um reduzido número de documentos. Portanto, termos com elevada frequência são considerados fracos para a discriminação de documentos. A segunda razão relaciona-se com a noção de conteúdo da informação. Termos que aparecem em vários contextos, tais como os artigos e preposições, geralmente não são considerados como referências para o contéudo. Esses termos não definem um tópico ou sub-tópico de um documento. Quanto mais documentos contém um termo (e mais contextos nos quais esse termo é empregado), então menor é a chance desse termo significar uma boa referência de contéudo. Consequentemente, é menos provável que esse termo contribua para a avaliação de relevância do usuário. Portanto, termos que aparecem em vários documentos são pouco utilizados pelos usuários na discriminação entre documentos relevantes e não-relevantes [Ruthven and Lalmas 2003].

Uma etapa comum do processo de indexação é remover todos os termos que aparecem frequentemente no conjunto de documentos e que não contribuem para a recuperação de conteúdo relevante (*stopword removal*, Figura 2 - c). Uma lista de termos que serão removidos é conhecida como *stop-list*. Essa lista pode ser genérica, isto é, uma lista que pode ser aplicada para a maioria dos conjuntos, ou uma lista especificamente criada para um conjunto individual [Ruthven and Lalmas 2003]. Faz-se necessário destacar que um termo não precisa obrigatoriamente aparecer na maioria dos documentos para que seja considerado um *stop term*. Por exemplo, [Crestani et al. 1995] menciona que a remoção

de todos os termos que apareceram em mais de 5% dos documentos não reduziu significantemente o desempenho de um sistema padrão de recuperação de informações.

Termos podem aparecer como variações linguísticas de uma mesma palavra. Por exemplo, na Figura 2, os termos *queries* e *query* constituem a forma plural e singular de um mesmo objeto. Ainda, os termos *expansion* e *expand* referem-se fundamentalmente à mesma atividade. Como a maioria dos sistemas de RI se baseiam em funções de casamento (*match*) de termos para recuperar documentos, esse tipo de variação no uso das palavras pode ser inconveniente para a busca do usuário. Por exemplo, se o usuário emite uma consulta com os termos 'hill walks', então um sistema de RI recuperará todos os documentos que contenham o termo 'walks', mas não os documentos contendo os termos 'hill walking', 'hill walk' ou 'hill walker', sendo que esses documentos poderiam conter informações relevantes para o usuário. De forma a evitar que o usuário tenha de instanciar todas as possíveis variações de cada termo empregado na consulta, muitos sistemas de indexação reduzem os termos para a sua forma raiz. Esse processo é conhecido como *stemming* (Stemming - Figura 2 - d) [Ruthven and Lalmas 2003].

Depois de passar por todas essas etapas, o documento resultante conterá um conjunto de palavras diferente do seu conjunto original. Para destacar essa diferença, essas palavras são chamadas de *termos*. Esses termos podem ser considerados como sendo a representação formal do documento. A coleção de termos de todo o conjunto de documentos é chamado de *corpus* [Hand 2007].

Quando o conjunto de documentos pesquisados não é muito grande, uma varredura diretamente no conteúdo dos documentos pode ser aplicada sem problemas. Entretanto, devido ao alto custo computacional, essa mesma abordagem não pode ser aplicada em grandes conjuntos de documentos, tal como é o ambiente Web. Para esse propósito, utiliza-se o chamado *índice invertido*, o qual é bem conhecido em RI. A ideia é trocar os papéis dos IDs dos documentos e dos seus termos. Desse modo, ao invés de acessar os documentos por seus IDs e então varrer seus conteúdos em busca de um termo específico, os termos que os documentos contém serão usados como suas chaves de acesso. A forma mais simples de um índice invertido é por meio de uma matriz relacionando termos e documentos, onde o acesso é feito através dos termos [Hand 2007]. Na sua representação booleana, cada célula da matriz contém 1 se o termo aparece no documento em questão, ou 0 caso contrário.

5. Ranking

A maioria dos sistemas de RI eram sistemas booleanos que permitiam ao usuário especificar sua necessidade de informação através de uma complexa combinação de operadores booleanos como *AND*, *OR* e *NOT*. Sistemas booleanos possuem desvantagens: não há nenhuma noção para classificação (*ranking*) de documentos e também é difícil para um usuário formular uma boa requisição para a busca [Singhal 2001].

Portanto, são necessárias informações adicionais sobre os termos, tais como quantidade, posicionamento, e outras informações de contexto. Uma abordagem direta é incorporar a quantidade de termos (frequências), conforme o modelo *TFIDF* (*Term Frequency Inverse Document Frequency*), o qual é amplamente conhecido no campo de RI e em buscas na web [Hand 2007].

5.1. Modelo Espaço Vetorial

A maioria dos sistemas de RI atribuem uma pontuação numérica (score) para cada documento recuperado e usam essa pontuação para classificar esses documentos. Vários modelos foram propostos para esse processo, sendo o modelo espaço vetorial o mais utilizado.

O modelo espaço vetorial define os documentos como sendo vetores (ou pontos) em um espaço Euclidiano multidimensional onde os eixos (dimensões) são representadas pelos termos dos documentos [Hand 2007].

Supondo que existam n documentos d_1, d_2, $\cdots$, d_n e m termos t_1, t_2, $\cdots$, t_m. Seja n_{ij} a quantidade de vezes que o termo t_i aparece no documento d_j. Considerando então a representação booleana, um documento d_j é representado como um vetor de m coordenadas $\vec{d} = (d_j^1 d_j^2 \cdots d_j^m)$, onde

$$d_j^i = \begin{cases} 0 & se \quad n_{ij} = 0 \\ 1 & se \quad n_{ij} > 0 \end{cases}$$

Na representação com frequência de termos (TF), as coordenadas do vetor de um documento são representadas em função de sua quantidade de termos. Para cada termo t_i e cada documento d_j, uma função $TF(t_i, d_j)$ é calculada. Essa função pode ser definida de várias formas, como por exemplo:

- Definição que considera a soma do número de ocorrências de cada termo do documento:

$$TF(t_i, d_j) = \begin{cases} 0 & se \quad n_{ij} = 0 \\ \dfrac{n_{ij}}{\sum_{k=1}^{m} n_{kj}} & se \quad n_{ij} > 0 \end{cases}$$

- Definição que considera o número máximo de ocorrências entre todos os termos do documento:

$$TF(t_i, d_j) = \begin{cases} 0 & se \quad n_{ij} = 0 \\ \dfrac{n_{ij}}{max_k . n_{kj}} & se \quad n_{ij} > 0 \end{cases}$$

Na representação com *frequência inversa do documento* (IDF), a ideia básica é reduzir o valor das coordenadas que correspondem aos termos que ocorrem em muitos documentos.

Para cada termo t_i, a sua medida IDF é calculada como uma proporção de documentos onde t_i ocorre em relação ao número total de documentos da coleção.

Tendo-se $\bigcup_1^n d_j$ a coleção de documentos e D_{ti} o conjunto de documentos onde o termo t_i aparece, assim como na abordagem TF, a medida IDF pode ser calculada de várias maneiras [Hand 2007].

Por exemplo, pode ser usada uma simples fração $|D|/|D_{ti}|$ ou mesmo uma função logarítmica, tal como:

$$\log \frac{1 + |D|}{|D_{ti}|}$$

Já na representação TFIDF, cada coordenada do documento vetorizado é obtida através do produto de suas componentes TF e IDF:

$$d_j^i = TF(t_i, d_j)IDF(t_i)$$

No modelo booleano, os termos da consulta são simplesmente casados contra a representação vetorial dos documentos, e os documentos que casam exatamente com essa consulta são retornados. O casamento exato não é possível no modelo TFIDF. Dessa forma, faz-se necessária alguma medida de proximidade entre a consulta e os documentos do conjunto. A ideia básica é representar a consulta também como um vetor no mesmo espaço vetorial do documento e então usar as propriedades métricas de espaços vetoriais. Para essa finalidade, as palavras-chave da consulta são primeiramente consideradas como sendo um documento.

Dado um vetor da consulta $\vec{q}$ e os vetores dos documentos $\vec{d_j}, j = 1, 2, \cdots, 20$, o objetivo de um mecanismo de busca é ordenar (*rank*) os documentos de acordo com as suas proximidades em relação a $\vec{q}$ [Hand 2007].

Há várias abordagems para este tipo de *ranking*. Uma opção é usar a *norma Euclidiana da diferença de vetores* $||\vec{q} - \vec{d_j}||$, definida como

$$||\vec{q} - \vec{d_j}|| = \sqrt{\sum_{i=1}^{m} (q^i - d_j^i)^2}$$

Essa medida é, de fato, a *distância Euclidiana* entre os vetores considerados como pontos no espaço Euclidiano.

A outra abordagem é usar o cosseno do ângulo entre o vetor da consulta e os vetores dos documentos. Quando os vetores são normalizados, essa medida é equivalente ao *produto escalar* $\vec{q}.\vec{d_j}$ definido como

$$\vec{q}.\vec{d_j} = \sum_{i=1}^{m} q^i d_j^i$$

Essa medida também é conhecida como *similaridade dos cossenos*.

6. Conclusão

O advento dos computadores tornou possível o armazenamento de grandes quantidades de informações. Assim, surgiu a necessidade de filtrar essas informações de forma a se obter apenas aquelas que sejam úteis para um determinado propósito. O campo da Recuperação de Informação nasceu justamente para atender a essa necessidade, possibilitando a descoberta da informação de uma forma mais rápida e fácil.

Técnicas desenvolvidas nesse campo tem sido empregadas em várias outras áreas e contribuíram para o surgimento de novas tecnologias que são utilizadas por pessoas

diariamente, tais como os mecanismos de busca na web. De fato, é seguro dizer que a busca na Web é a aplicação mais importante da RI.

Com o crescimento exponencial do volume de informações disponíveis, a área de Recuperação da Informação certamente desempenhará um papel cada vez mais importante no futuro, recuperando o que é útil e descartando o que não é.

Como trabalho futuro, a técnica de feedback de relevância será estudada para compreender como essa técnica pode melhorar ainda mais os resultados retornados por um mecanismo de busca.

Referências

Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

Crestani, F., Ruthven, I., Sanderson, M., and van Rijsbergen, C. (1995). The troubles with using a logical model of ir on a large collection of documents.

Hand, D. J. (2007). Data mining the web: Uncovering patterns in web content, structure, and usage by zdravko markov, daniel t. larose. *International Statistical Review*, 75(3):409–409.

Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(1).

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.