Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

META-MODELO FUNCIONAL PARA RECUPERAÇÃO DE INFORMAÇÃO

Luciene Chagas de Oliveira

Uberlândia - MG
Fevereiro de 2006

META-MODELO FUNCIONAL PARA RECUPERAÇÃO DE INFORMAÇÃO

Por

Luciene Chagas de Oliveira

DISSERTAÇÃO APRESENTADA À UNIVERSIDADE FEDERAL DE UBERLÂNDIA, MINAS GERAIS, COMO PARTE DOS REQUISITOS EXIGIDOS PARA OBTENÇÃO DO TÍTULO DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO

Área de concentração: Banco de Dados.

Orientador: Ilmério Reis da Silva - UFU
Co-Orientrador: João Nunes de Souza - UFU

FEVEREIRO DE 2006

FICHA CATALOGRÁFICA
Elaborado pelo Sistema de Bibliotecas da UFU / Setor de Catalogação e Classificação

O48m	Oliveira, Luciene Chagas de, 1980- Meta-Modelo funcional para recuperação de informação / Luciene Chagas de Oliveira. - Uberlândia, 2006. 90f. : il. Orientador: Ilmério Reis da Silva. Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Ciência da Computação. Inclui bibliografia. 1. Recuperação da informação - Teses. 2. Banco de Dados - Teses. 3. Algoritmos de computador - Teses. I. Silva, Ilmério Reis da. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Ciência da Computação. III. Título. CDU: 681.3.07

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada "**Meta-Modelo Funcional para Recuperação de Informação**" por **Luciene Chagas de Oliveira** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

Uberlândia, 22 de fevereiro de 2006

Orientador: _____
Prof. Dr. Ilmério Reis da Silva
Universidade Federal de Uberlândia UFU / MG

Co-Orientador: _____
Prof. Dr. João Nunes de Souza
Universidade Federal de Uberlândia UFU / MG

Banca Examinadora: _____
Prof^{a}. Dr^{a}. Ana Paula Laboissière Ambrósio
Universidade Federal de Goiás UFG / GO

_____
Prof. Dr. Sérgio de Mello Schneider
Universidade Federal de Uberlândia UFU / MG

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Data: Fevereiro, 2006

Autora:	**Luciene Chagas de Oliveira**
Título:	**Meta-Modelo Funcional para Recuperação de Informação**
Faculdade:	**Faculdade de Computação**
Grau:	**Mestrado**

_____

Autora

Dedicatória

Aos meus pais Jair e Terezinha, aos meus irmãos Eduardo e Liliane e ao meu namorado Wellington

Agradecimentos

Primeiramente, agradeço a Deus, acima de tudo, que permitiu este momento de grande importância em minha vida.

Ao meu orientador e ao meu co-orientador, Profs. Drs. Ilmério R. Silva e João N. Souza, sou profundamente grata pela contínua orientação, pela confiança, pela amizade, por suas idéias e pelos conselhos durante todo o desenvolvimento deste trabalho.

Aos membros da banca, Prof. Dr. Ana Paula L. Ambrósio e Prof. Dr. Sérgio de Mello Schneider pela colaboração.

Ao Prof. Edleno Moura da UFAM pelas discussões teóricas e sugestões para o meu trabalho.

À toda minha família, pelo incentivo durante esta jornada, especialmente aos meus pais Jair e Terezinha, aos meus irmãos Eduardo e Liliane, e ao meu namorado Wellington pelo amor e carinho incondicional, por nortearem meus caminhos e por me darem forças para enfrentar os desafios da vida.

À todos os professores e amigos da Pós Graduação da Universidade Federal de Uberlândia, sou grata pelos ensinamentos, companheirismo e pela amizade. Agradeço, em especial, os colegas do grupo de Recuperação de Informação da UFU, Juliana Franciscani e Daniel Gonzaga, que estiveram presentes em alguns momentos importantes na elaboração deste trabalho e aos meus colegas da Pós Graduação, Daniel Furtado, Elaine Ribeiro, Paulo Vidica, Rogério Novo e Fábio Divino, pela amizade, conselhos, pelos trabalhos e estudos que realizamos juntos durante as disciplinas cursadas no mestrado.

Aos colegas e amigos da SWB pela amizade e apoio durante a execução deste trabalho.

Finalmente, agradeço a todos que contribuiram de alguma forma para a conclusão deste trabalho.

Resumo

Modelagem é uma das tarefas centrais no desenvolvimento de sistemas de recuperação de informação. Uma ferramenta para modelagem muito utilizada para desenvolvimento de um novo modelo de recuperação de informação é um *framework* genérico. Estes *frameworks* podem ser vistos como meta-modelos formais que possibilitam formalmente descrever e investigar a semântica do processo de recuperação e tornam possível o raciocínio sobre as características e propriedades de modelos de recuperação de informação (RI). Com o crescimento e as diferenças entre as estratégias e modelos de RI, a modelagem formal vem se tornando cada vez mais importante.

Nesta dissertação, propomos um *framework* genérico e formal para definição de modelos de RI chamado de Estrutura Funcional. Este *framework* é um meta-modelo para modelos de RI que define um nível de abstração permitindo a representação, combinação, formulação e comparação de equivalência entre modelos de RI. Com este meta-modelo, modelos de RI podem ser representados em uma única linguagem comum, tornando mais fácil o estudo de características e propriedades dos modelos e a combinação desses modelos. O *framework* também fornece um formalismo que permite a comparação de modelos sem a necessidade de realizar experimentos.

Além disso, mostramos aqui exemplos de como representar os modelos clássicos de RI e construímos um modelo baseado em distância equivalente ao modelo clássico vetorial usando a estrutura funcional. Também analisamos a combinação de múltiplas evidências, apresentamos dois estudos de caso do uso da estrutura funcional para combinar múltiplas evidências nos contextos de redes bayesianas e modelo de espaço vetorial. Mostramos que a combinação de múltiplas evidências na rede bayesiana de crença pode ser realizada de várias formas, sendo que cada uma corresponde à uma função de similaridade no modelo vetorial. A análise dessa correspondência é feita através da estrutura funcional. Com isso, mostramos que o *framework* permite-nos desenvolver novos modelos e ajuda desenvolvedores a modificar esses modelos para extendê-los com novas fontes de evidências. Como aplicação do meta-modelo funcional, apresentamos também as idéias de desenvolvimento de uma meta-ferramenta para comparação experimental entre modelos de RI.

Palavras-chave: Estrutura Funcional, Modelos de Recuperação de Informação, Modelos Formais, Meta-Modelo, Combinação de Múltiplas Evidências

Abstract

Modelling is one of the central tasks in the development of information retrieval systems. A useful tool for developing a new information retrieval model is a generic framework. This frameworks can be seen as formal meta-models that make possible to describe and to investigate formally the semantics of the retrieval process and becomes possible to reason about features and properties of information retrieval models (IR). With the growth and the differences between the IR strategies and models formal modelling comes becoming more and more important.

In this dissertation, we propose a generic and formal framework for defining IR models named Functional Framework. This framework is a meta-model for IR models, defining a level of abstraction that allows the representation, formulation and comparison of IR models. With this meta-model, IR models can be represented in a unique common language, which makes the study of characteristics and properties of the models and the combination of these models easier. The framework also provides a formalism that permits the comparison of models without the need to carry out experiments.

Moreover, we show examples of how to represent the three classic IR models and we design a model based on distance equivalent to the classic vector model using the framework functional. We also analyze the combination of multiple evidence, presenting two case studies of the use of the framework to combine multiple evidence in contexts bayesian belief networks and in the vector space model. We show that the combination of multiple evidence in the bayesian belief network can be carried at in of several ways, being that each form corresponds to a similarity function in the vector model. The analysis of this correspondence is made through the functional framework. We show that the framework allows us to design new models and helps designers to modify these models to extend them with new evidence sources. As application of the functional meta-model, we also present the ideas of development of a meta-tool for experimental comparison between IR models.

Keywords: Functional Framework, Information Retrieval Models, Formal Models, Meta-Model, Combination of Multiple Evidence

Sumário

Lista de Figuras **xv**

Lista de Acrônimos **xvii**

Lista de Símbolos **xix**

1 Introdução **1**

1.1 Recuperação de Informação 1

1.2 Meta-Modelos Formais em Recuperação de Informação 3

1.3 Objetivos e Contribuições 4

1.4 Organização da Dissertação 6

2 Fundamentos de RI **9**

2.1 Modelos Clássicos 9

2.1.1 Modelo Booleano 10

2.1.2 Modelo Clássico Vetorial 11

2.1.3 Modelo Probabilístico 12

2.2 Redes Bayesianas 14

2.2.1 Conceitos Básicos 15

2.2.2 O Modelo de Redes de Crença para RI 21

2.2.3 Rede de Crença para o Modelo Clássico Vetorial 23

2.3 Fontes de Evidências: Análise de *Links* 24

2.4 Meta-Modelos Alternativas 26

2.4.1 Meta-Modelos Algébricos 26
2.4.2 Meta-Modelos Baseados em Lógica 30

3 Trabalhos Relacionados 33
3.1 Meta-Modelos Formais 33
3.1.1 Meta-Modelos Algébricas 33
3.1.2 Meta-Modelos Baseados em Probabilidades 34
3.1.3 Meta-Modelos Baseados em Lógica 35
3.2 Combinando Informação de Evidências para RI 36

4 Estrutura Funcional para RI 39
4.1 Fundamentos da Estrutura Funcional 39
4.1.1 Representação de Modelos 40
4.1.2 Comparação de Modelos 43
4.2 Representação dos Modelos Clássicos 45
4.2.1 Modelo Funcional Vetorial 46
4.2.2 Modelo Funcional Booleano 47
4.2.3 Modelo Funcional Probabilístico 48
4.3 Comparação e Construção de Modelos usando a Estrutura Funcional 50
4.3.1 Modelo Baseado em Distância 50
4.3.2 Modelo Funcional Baseado em Distância 52
4.3.3 Equivalência entre os Modelos Funcionais que representam o Modelo Vetorial e o Modelo Baseado em Distância 53

5 Combinação de Múltiplas Evidências usando a Estrutura Funcional 57
5.1 Estudo de Caso 1: Do Modelo de Redes Bayesianas para o Modelo Vetorial 57
5.1.1 Modelo de Redes de Crença para Combinar Múltiplas Fontes de Evidências . 58
5.1.2 Modelo Funcional de Redes de Crença para Combinar Múltiplas Fontes de Evidências 60
5.1.3 Modelo Vetorial para Combinar Múltiplas Fontes de Evidências 62
5.1.4 Modelo Funcional Vetorial para Combinar Múltiplas Fontes de Evidências . 62

5.2 Estudo de Caso 2: Do Modelo Vetorial para o Modelo de Redes Bayesianas 64
5.2.1 Modelo Vetorial para Combinar Múltiplas Fontes de Evidências 64
5.2.2 Modelo Vetorial Funcional para Combinar Múltiplas Fontes de Evidências . 65
5.2.3 Modelo de Redes de Crença para Combinar Múltiplas Fontes de Evidências . 65
5.2.4 Modelo Funcional de Redes de Crença que Combina Múltiplas Fontes de Evidências 66

6 Proposta de uma Ferramenta para Avaliação de Desempenho de SRI 69
6.1 Introdução 69
6.2 Especificação dos Modelos Funcionais 71
6.3 Especificação da Coleção de Referência 72
6.4 Especificação da Medida de Avaliação 73
6.4.1 Precisão 74
6.4.2 Revocação 75
6.4.3 Precisão nos X primeiros 75
6.4.4 Precisão-R 75
6.4.5 Medida-E 75
6.5 Processo de Recuperação 76
6.6 Processo de Comparação Relativa ou Avaliação de Resultados 77

7 Conclusões e Trabalhos Futuros 79
7.1 Conclusões 79
7.2 Trabalhos Futuros 80

Referências bibliográficas 82

Lista de Figuras

2.1 Conjuntos para Representação da Regra do Produto da Teoria da Probabilidade . . . 17
2.2 Interpretação gráfica para a Regra da Probabilidade Total 18
2.3 Nós pais de um nó em uma Rede Bayesiana 20
2.4 Exemplo de uma Rede Bayesiana 21
2.5 Rede Bayesiana para uma consulta q composta pelos termos k_1 e k_i 22
2.6 Conjunto de *Hubs* e Autoridades 26

4.1 Representação do mapeamento entre os modelos Ψ_a e Ψ_b da definição de equivalência 44
4.2 Esquema geral para comparação de equivalência entre modelos de RI 45
4.3 Representação do documento $\vec{d_j}$ e da consulta $\vec{q}$ em um espaço bi-dimensional e seus vetores normalizados $\vec{d'_j}$ e $\vec{q'}$ 51
4.4 Representação dos documentos $d\vec{f}_j$ e $d\vec{f}_k$, da consulta $q\vec{f}$ em um espaço bi-dimensional , seus vetores normalizados, $d\vec{f}'_j$, $d\vec{f}'_k$ e $q\vec{f}'$, e a distância entre eles 53

5.1 Modelo de rede de crença para combinar múltiplas fontes de evidências 58
5.2 Modelo genérico vetorial para combinação de múltiplas fontes de evidências 63
5.3 Modelo genérico de redes de crença para combinar múltiplas fontes de evidências . . 66

6.1 Representação gráfica da proposta da ferramenta para avaliação de desempenho entre modelos de RI 70
6.2 Conjuntos para definição de precisão e revocação 74

Lista de Acrônimos

CACM - Communications of the ACM

CFC - Cystic Fibrosis Collection

CISI - Collection of Institute of Scientific Information

IDF - Inverse Document Frequency

RI - Recuperação de Informação

SRI - Sistema de Recuperação de Informação

TREC - Text Retrieval Conferences

WWW - World Wide Web

Lista de Símbolos

N	-	Número de documentos na coleção
t	-	Número de termos na coleção ou tamanho do vocabulário
d_j	-	j-ésimo documento da coleção
k_i	-	i-ésimo termo de um documento
q	-	Consulta do usuário
$w_{i,j}$	-	Função peso do termo k_i no documento d_j
$w_{i,q}$	-	Função peso do termo k_i no documento q
$freq_{i,j}$	-	Frequência natural do termo k_i no documento d_j, isto é, número de vezes que k_i aparece em d_j
$IDF(k_i)$	-	Freqüência inversa de documentos do termo k_i em uma coleção
$sim(d_j, q)$	-	Função similaridade entre o documento d_j e a consulta q
$\mathbf{R}_q$	-	Conjunto de documentos relevantes para a consulta q
$\overline{\mathbf{R}_q}$	-	O complemento de $\mathbf{R}_q$
$P(X\|Y)$	-	Probabilidade de X dado Y.
$P(k_i\|\mathbf{R}_q)$	-	Probabilidade de um termo k_i estar presente em um documento escolhido aleatoriamente do conjunto $\mathbf{R}_q$
$P(\mathbf{R}_q\|d_j)$	-	Probabilidade do documento d_j ser relevante para a consulta q no conjunto $\mathbf{R}$
df_j	-	j-ésimo documento funcional de uma coleção na estrutura funcional
qf	-	Consulta funcional do usuário na estrutura funcional
g_j	-	Função peso de um termo no documento d_j na estrutura funcional
g_q	-	Função peso de um termo no documento q na estrutura funcional
$D(\vec{d_j}, \vec{q})$	-	Distância de *Minkowski* entre os vetores $\vec{d_j}$ e $\vec{q}$
$R_{j,q}$	-	Função de *ranking* ou similaridade calculada pelo modelo vetorial
$E_{i,j}$	-	Valor da i-ésima evidência em relação ao documento d_j
$E_{i,q}$	-	Valor da i-ésima evidência em relação a consulta q
$\mathbf{O}_f$	-	Conjunto de objetos funcionais na estrutura funcional
$\mathbf{D}_f$	-	Conjunto de documentos funcionais
$\mathbf{Q}_f$	-	Conjunto de consultas funcionais
$\mathbf{T}_f$	-	Conjunto de termos funcionais dos documentos e consultas funcionais
$\mathbf{C}_f$	-	Coleção de referência funcional
Ψ	-	Modelo funcional
$\Delta(of_j, of_i)$	-	Função similaridade entre dois objetos funcionais of_j e of_i na estrutura funcional
η	-	Constante de normalização
β	-	Define a relativa importância de precisão e revocação para a medida-E

Capítulo 1

Introdução

A área de Recuperação de Informação (RI) possui grande importância em Ciência da Computação há várias décadas e tem experimentado um maior interesse da comunidade científica devido à grande disponibilidade de documentos existentes hoje na forma digital, principalmente na Web. Recuperação de informação estuda o armazenamento e a recuperação automática de documentos. A modelagem é um dos tópicos de pesquisa centrais e mais ativos em RI. Neste trabalho propomos um meta-modelo, chamado Estrutura Funcional, como uma ferramenta para ajudar projetistas na tarefa de desenvolvimento, representação e comparação de modelos de RI. Neste capítulo, apresentamos alguns conceitos de recuperação de informação, meta-modelos formais em RI, discutimos os objetivos e as contribuições de nosso trabalho.

1.1 Recuperação de Informação

Recuperação de Informação (RI) é uma área que estuda o armazenamento, classificação, agrupamento e recuperação automática de documentos. A abundância de informações na Web é uma das principais razões para sua crescente popularidade. A facilidade do uso e acessibilidade da Web fez dela uma ferramenta muito importante, não somente para comunicação, mas também para armazenamento e compartilhamento de informação. Do ponto de vista de RI, a Web pode ser vista como grande repositório de dados contendo documentos, ou páginas Web, que são interconectados.

O problema central em recuperação de informação é encontrar informações de interesse dos

usuários. A principal ferramenta usada para resolver este problema é o emprego de sistemas de recuperação de informação (SRI). A área de recuperação de informação (RI) apresentou importantes resultados, desde o seu início, nas tarefas de localizar e classificar documentos em sistemas bibliográficos, até então restritos a bibliotecas e redes de menor escala. Mais recentemente, o emprego de RI para busca de informações na Web contribuiu enormemente para a criação de máquinas de buscas.

O usuário de um SRI ou de máquinas de buscas da Web geralmente traduz sua necessidade de informação mediante a especificação de uma consulta. Tradicionalmente, essa consulta é um conjunto de palavras-chaves (chamadas termos) que são usadas para recuperar documentos em uma coleção. Assim, o SRI apresenta os documentos em ordem descrescente de relevância que satisfazem a consulta submetida. A noção de relevância é um conceito fundamental em RI [42], indica a importância de um documento para uma consulta sendo um componente chave para determinar o *ranking* de documentos.

A Web tem algumas características únicas, a maioria dos dados armazenados são muito volatéis, com documentos constantemente sendo modificados, removidos ou adicionados. Alguns dados também contém informações errôneas ou ruídos. Os documentos não contém apenas textos, mas também sons, vídeos, imagens e outros tipos de mídias. Além disso, a maioria dos usuários da Web não são especializados, com poucas habilidades no uso de sistemas de recuperação de informação, e também têm interesses diversificados. As consultas geralmente são vagas [60, 61]. Neste cenário, a recuperação baseada em texto torna-se insuficiente. Novas fontes de evidências [1] têm sido utilizadas para permitir determinar com maior precisão a relevância de documentos que satisfazem a consulta do usuário. Por exemplo: informações extraídas de imagens e outros tipos de mídias podem ser combinados com a análise de texto; informações da estrutura interna de documentos da Web podem ser usadas para indicar a importância do conteúdo dos documentos e o uso de *log* de usuários pode ser analisado para indicar preferências.

Os sistemas de recuperação de informação são construídos baseados em modelos de RI. Diversos modelos de recuperação têm sido estudados e propostos em RI. Dentre eles estão os modelos de espaço vetorial [1, 23, 26, 53], modelos probabilísticos [15, 25, 38, 64] e modelos baseados em lógica [24, 3, 66].

[1] Evidência é uma informação adicional sobre um fato

Os tipos de modelos de RI desenvolvidos podem ser divididos em:

- Modelos clássicos de RI: existem três modelos clássicos de RI (booleano, vetorial e probabilístico) [1];
- Modelos alternativos: por exemplo, técnicas baseadas em conjuntos *fuzzy*, booleano generalizado, vetorial generalizado, entre outros [1, 2];
- Modelos lógicos de RI [9, 10, 12];
- Interação de RI [7, 21, 17];
- Inteligência Artificial: existem modelos baseados em conhecimento, em redes neurais artificiais, em algoritmos genéticos e em linguagem de processamento natural [13, 14, 34, 29];

Os diferentes tipos de modelos de RI refletem a complexidade de RI em geral e da modelagem em RI. Os tipos de modelos de RI diferem, em sua maioria, na forma que os objetos (documentos, imagens,etc) são representados e como a recuperação é definida. Os modelos algébricos ou matemáticos geralmente representam os objetos como uma seqüência de números (tradicionalmente chamados de vetores), e define a recuperação como um relacionamento entre os números. A Lógica em RI assume os objetos como representações (exemplo, coleções de sentenças) e a recuperação como uma inferência lógica. A interação em RI visualiza os objetos como elementos interconectados e a recuperação como memórias de elementos chamados por uma consulta. A Inteligência Artificial para RI visualiza os objetos como conhecimento, e a recuperação como alguma razão ou como neurônios ou como regiões de ativação.

1.2 Meta-Modelos Formais em Recuperação de Informação

A implementação de qualquer nova idéia para melhorar a qualidade do *ranking* e a exatidão de um sistema de recuperação da informação requer geralmente uma primeira etapa de modelagem. A modelagem é uma tarefa complexa, e também importante, em sistemas de recuperação modernos, tais como máquinas de busca na Web e sistemas de busca tradicionais. Nestes casos é comum termos mais que uma fonte de evidência avaliada para ser explorada pelo modelo na tarefa de fornecer respostas

para uma dada consulta. Esta abundância de fontes de evidência certamente oferece uma oportunidade para o desenvolvimento de sistemas mais eficientes, mas também é um desafio para desenvolvedores de um modelo de RI.

Uma ferramenta muito utilizada para desenvolvimento de um novo modelo de recuperação de informação são os *frameworks* genéricos. Os *frameworks* genéricos podem ser vistos como meta-modelos formais que ajudam desenvolvedores na produção de novos modelos de recuperação de informação. Estes meta-modelos são utilizados não somente para projetar novos modelos, mas também para representação de modelos anteriormente propostos facilitando o estudo de propriedades e características dos modelos de RI e para ajudar desenvolvedores a modificar esses modelos ou extendê-los com novas evidências de forma que eles se tornem mais eficientes e flexíveis.

Um formalismo em RI geralmente utiliza notação matemática para a representação de estratégias de RI ou modelos de RI. Um meta-modelo formal consiste na notação utilizada para detalhar um modelo de RI para formalmente estudar suas propriedades e características [11, 46]. A representação em notação matemática de estratégias de recuperação é um importante assunto de pesquisa em RI. Estes meta-modelos permitem que diferentes características dos modelos possam ser combinadas em um mesmo plano de representação.

1.3 Objetivos e Contribuições

Este trabalho apresenta um meta-modelo formal baseado em funções denominado estrutura funcional, cujo objetivo é ajudar desenvolvedores na tarefa de desenvolvimento de modelos de RI, permitindo a representação, combinação, construção e comparação de modelos de RI.

As principais contribuições desta dissertação são:

- A proposta de um meta-modelo funcional com os seguintes objetivos:
  - ser um unificador de modelos;
  - fornecer expressividade, capacidade de abstração e identificação de componentes e relações relevantes em problemas de RI;
  - simplificar a construção de modelos através de funções;

  - ser capaz de representar além dos modelos clássicos, modelos que combinam evidências e o conjunto de modelos que podem ser expressos por meio de algoritmos;
  - comparar modelos quanto a sua similaridade (equivalência) ou não, sem realizar experimentos;

- A construção de quatro novos modelos de recuperação de informação utilizando a estrutura funcional:
  - modelo baseado em distância;
  - modelo vetorial para combinar múltiplas evidências utilizando a função de similaridade noisy-OR do contexto de redes bayesianas;
  - modelo vetorial para combinar múltiplas evidências utilizando a função de similaridade cosseno;
  - modelo de redes bayesianas para combinar múltiplas evidências utilizando a função de similaridade cosseno;

- A proposta de uma ferramenta para avaliação experimental de modelos de RI;

Nosso objetivo é apresentar a estrutura funcional como um meta-modelo para RI e suas aplicações. A área de teorias para RI tem uma orientação pragmática muito forte refletida nos interesses comerciais e na ênfase de avaliações formais no ambiente acadêmico [48]. Através do meta-modelo funcional buscamos o desenvolvimento de uma teoria para RI que engloba muitos aspectos diferentes de recuperação.

Dentre as aplicações do meta-modelo funcional temos a criação de modelos de RI, a comparação de equivalência sem a necessidade de realizar experimentos e a comparação relativa utilizando experimentos através da ferramenta que apresentamos neste trabalho. Outra aplicação é a comparação algébrica ou relativa sem realizar experimentos que propomos como pesquisa futura. Com este meta-modelo, modelos de RI podem ser representados em uma linguagem comum, tornando mais fácil o estudo de características e propriedades dos modelos, a combinação desses modelos e a comparação relativa entre modelos utilizando experimentos que mostramos através de uma proposta de uma fer-

ramenta para avaliação de resultados. Assim, através da linguagem funcional podemos pensar nos modelos de RI em um nível mais alto de abstração.

A principal motivação de nosso trabalho é a construção de um novo meta-modelo que por ser baseado em funções se diferencia dos outros meta-modelos propostos na literatura, a saber, baseados em lógica, baseados em probabilidade e outros meta-modelos algébricos. Alguns destes meta-modelos serão brevemente descritos no Capítulo 2. A estrutura funcional não é tão limitada quanto os meta-modelos probabilísticos que são difíceis de serem aplicadas em alguns contextos e nem tão abstrata quanto os meta-modelos lógicos que carecem de exemplos de aplicação. O meta-modelo funcional proposto é prático no sentido de implementação dos problemas de RI e não tão genérico permitindo trabalhar com aplicações teóricas e práticas.

Um trabalho relacionado à dissertação foi publicado com o título *Comparison of Representations of Multiple Evidence using a Functional Framework for IR* [56].

1.4 Organização da Dissertação

O conteúdo desta dissertação está organizado em 7 capítulos, como descrito a seguir.

No Capítulo 2 introduzimos alguns fundamentos e conceitos básicos em Recuperação de Informação. Discutimos os modelos clássicos para recuperação de informação e os modelos de redes bayesianas, apresentamos os conceitos sobre análise de *links* e descrevemos alguns meta-modelos alternativos.

No Capítulo 3 abordamos os principais trabalhos relacionados à estrutura funcional e à combinação de evidências em modelos de RI.

Formalizamos os conceitos da estrutura funcional para RI, representamos os modelos clássicos na estrutura funcional, e projetamos um novo modelo baseado em distância equivalente ao modelo clássico vetorial no Capítulo 4.

No Capítulo 5, usamos a estrutura funcional para analisar dois estudos de caso para combinar múltiplas evidências nos modelos de redes bayesianas e vetorial.

No Capítulo 6 apresentamos uma proposta para construção de uma ferramenta para avaliação experimental de resultados utilizando o meta-modelo funcional.

Finalmente, no Capítulo 7 concluímos o trabalho, discutimos as vantagens potenciais da utilização da estrutura funcional e apresentamos algumas direções para pesquisas futuras.

Capítulo 2

Fundamentos de RI

Este capítulo introduz os fundamentos de RI para a proposta de um meta-modelo para RI e sua aplicação. Os modelos clássicos são descritos na Seção 2.1 e modelos de redes bayesianas são descritos na Seção 2.2, esses modelos serão posteriormente representados e utilizados como aplicação do meta-modelo funcional (Capítulos 4 e 5). Na Web, as fontes de evidências comumente utilizadas são os *links*. Alguns modelos discutidos nesta dissertação utilizam análise de *links*,isto é, a extração de informação da estrutura de *links* na Web. Então, alguns conceitos de análise de *links* são descritos na Seção 2.3. Finalmente, a Seção 2.4 apresenta alguns meta-modelos alternativos à estrutura funcional que motivaram este trabalho.

2.1 Modelos Clássicos

Um modelo de recuperação de informação representa documentos e consultas para predizer o que um usuário considera relevante para sua necessidade de informação. São três os modelos clássicos seguidos por sistemas de RI para determinar a relevância de documentos: booleano, vetorial e probabilístico.

Os modelos clássicos, utilizados no processo de recuperação de informação, apresentam estratégias de busca de documentos similares à consulta. Estes modelos consideram que cada documento é descrito por um conjunto de termos, considerados como mutuamente independentes. Associa-se a cada termo k_i e um documento d_j um peso $w_{i,j} \geq 0$, que quantifica o peso do termo k_i no docu-

mento d_j. Este peso reflete a importância do termo k_i no documento d_j. Analogamente a cada par termo-consulta (k_i, q) associa-se o peso $w_{i,q}$.

Modelos mais avançados têm sido propostos, mas ainda existe uma grande necessidade por novos arcabouços que permitam a melhoria na qualidade das respostas. Descrevemos abaixo os três modelos clássicos.

2.1.1 Modelo Booleano

O modelo booleano foi o primeiro modelo utilizado em RI e o mais utilizado até meados da década de 1990, apesar das alternativas que surgiram desde o final dos anos 1960 [40, 49, 67].

Este modelo considera uma consulta como uma expressão booleana convencional, que liga seus termos através de conectivos lógicos AND, OR e NOT. Nesse modelo um documento é considerado relevante ou irrelevante para uma consulta; não existe resultado parcial e não há informações que permitam a ordenação do resultado da consulta.

O fato de o modelo booleano não possibilitar a ordenação dos resultados por ordem de relevância é uma de suas principais desvantagens, já que esta classificação é uma característica considerada essencial em muitos dos sistemas de RI modernos, por exemplo, nas máquinas de busca.

Outra característica que pode ser considerada uma desvantagem no caso de usuários inexperientes é o uso de operadores booleanos. Para os usuários que conhecem álgebra booleana, os operadores podem ser considerados uma forma de controlar/direcionar o sistema. Se o conjunto resposta é pequeno ou grande, eles saberão quais operadores utilizar para produzir um conjunto de respostas maior ou menor. No entanto, para usuários comuns, o uso dos operadores booleanos não é intuitivo, pois é diferente do uso de suas palavras equivalentes em linguagem natural.

As vantagens do modelo booleano são a facilidade de implementação e a expressividade completa das expressões. Apesar dos problemas deste modelo, dada a sua simplicidade e seu formalismo, recebeu uma enorme atenção a alguns anos atrás e foi adotado por muitos sistemas bibliográficos comerciais. Além disso, existem variações deste modelo por exemplo, os modelos fuzzy e booleano estendido. Essas variações, em geral, mantêm a expressividade de consultas booleanas e trazem respostas ordenadas por relevância [1].

2.1.2 Modelo Clássico Vetorial

No modelo de espaço-vetorial, ou simplesmente modelo vetorial, os documentos e as consultas são representados por um vetor em um espaço de termos. O conjunto de termos de uma coleção de documentos é chamado de vocabulário. Cada termo possui um peso associado que indica seu grau de importância no documento. Em outras palavras, os documentos e as consultas possuem vetores associados a cada um.

Cada elemento do vetor de termos é considerado uma coordenada dimensional. Assim, os documentos e consultas do usuário são representados como vetores de termos em um espaço t-dimensional, onde t é o número de termos ou tamanho do vocabulário. O j-ésimo documento em uma coleção de documentos é denotado por d_j. Um termo é uma palavra que semanticamente ajuda a lembrar o tema principal do documento. Um termo é denotado por k_i. Então o vetor associado ao documento d_j é dado por $\vec{d_j} = (w_{i,j}, w_{2,j}, \ldots, w_{t,j})$, onde $w_{i,j}$ é o peso associado ao termo k_i no documento d_j. Consultas também são representadas por vetores. Uma consulta é um conjunto de termos que expressa a necessidade do usuário, e é denotada por q. O vetor associado à consulta q é $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$, onde $w_{i,q}$ é o peso associado ao termo k_i na consulta q.

Cada dimensão deste espaço é associada com um vetor de termos $\vec{k_i}$. Estes vetores de termos são ortogonais, ou seja, $i \neq j \Rightarrow \vec{k_i} \bullet \vec{k_j} = 0$. Isto indica que assumimos que termos ocorrem independentemente dentro dos documentos e consultas. Além disso, $|\vec{k_i}| = 1$.

O modelo vetorial propõe avaliar o grau de similaridade entre um documento d_j e uma consulta q como uma correlação entre vetores $\vec{d_j}$ e $\vec{q}$. Esta correlação pode ser quantificada pelo cosseno do ângulo entre estes vetores. Então, a *fórmula de similaridade* é definida como:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2.1}$$

Os pesos $w_{i,j}$ e $w_{i,q}$ quantificam a importância do termo k_i para a consulta e para os documentos, respectivamente. Os pesos podem ser calculados de diferentes maneiras [71]. Uma delas é mostrado como se segue.

Seja N o número total de documentos na coleção e n_i o número de documentos em que o termo k_i aparece. Seja $freq_{i,j}$ a frequência natural do termo k_i no documento d_j, isto é, o número de vezes

que o termo k_i é mencionado no texto do documento d_j. Se o termo k_i não aparece no documento d_j, então $freq_{i,j}$ = 0. Cada frequência do termo fornece uma medida de como o termo descreve o conteúdo do documento, denominada *caracterização intra-documento*.

Para cada termo é calculado também a frequência inversa dos documentos onde o termo aparece, IDF, que fornece uma *caracterização inter-documento*. A motivação para o seu uso é que termos que aparecem em muitos documentos não são úteis para distinguir um documento relevante de um documento não relevante. O $IDF(k_i)$, freqüência inversa de documentos do termo k_i em uma coleção é dado por:

$$IDF(k_i) = \log \frac{N}{n_i} \tag{2.2}$$

O peso do termo no documento é dado pela fórmula $w_{i,j} = freq_{i,j} \cdot IDF(k_i)$ e o peso do termo na consulta é dado por $w_{i,q} = freq_{i,q} \cdot IDF(k_i)$ [52].

A norma do documento d_j de uma coleção é dada por:

$$|\vec{d_j}| = \sqrt{\sum_{i=1}^{t} w_{i,j}^2} = \sqrt{\sum_{i=1}^{t} (freq_{i,j} \cdot IDF(k_i))^2} \tag{2.3}$$

Calculados os graus de similaridade pela Equação (2.1), é possível montar uma lista ordenada de todos os documentos ordenados por seus respectivos graus de relevância à consulta ou *ranking*. Um documento pode ser recuperado mesmo se ele satisfazer a consulta somente parcialmente. Assim, os documentos mais similares à consulta ficarão no topo desta ordenação.

Este é um modelo muito utilizado em sistemas de recuperação de informação. As principais razões para isto são a sua rapidez no processo de busca, a sua simplicidade, a flexível estratégia de agrupamento e a boa precisão na recuperação de documentos de coleções genéricas [1, 23, 53].

2.1.3 Modelo Probabilístico

No modelo probabilístico, os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos. Este modelo descreve documentos considerando pesos binários

que representam a presença ou ausência de termos. O *ranking* gerado por ele tem como base o cálculo da probabilidade de que um documento seja relevante para uma consulta. A principal ferramenta matemática do modelo probabilístico é o teorema de Bayes (veja detalhes em [65] e Seção 2.2).

É baseado no princípio da ordenação probabilística (*Probability Ranking Principle*): dada uma consulta q e um documento d_j de uma coleção, o modelo tenta estimar a probabilidade do usuário localizar o documento d_j relevante. O modelo assume que esta probabilidade de relevância depende somente das representações da consulta e do documento. O modelo probabilístico assume que há um conjunto ótimo de documentos, rotulados $\mathbf{R}_q$ que maximiza toda a probabilidade de relevância para o usuário. Documentos no conjunto $\mathbf{R}_q$ são considerados relevantes para uma consulta q. Documentos que não estão neste conjunto são considerados não relevantes.

Como não sabemos quais são os documentos relevantes e em uma situação prática, o conjunto relevante $\mathbf{R}_q$ deve ser inicialmente estimado e esperamos metlhorá-lo por meio de interações com o usuário.

Os pesos neste modelo são todos binários, isto é, $w_{i,j} \in \{0,1\}$ e $w_{i,q} \in \{0,1\}$. Seja o conjunto $\mathbf{R}_q$ o conjunto de documentos que foram estimados como relevantes para a consulta q, isto é, uma estimativa para o conjunto ideal, e seja $\overline{\mathbf{R}_q}$ o complemento de $\mathbf{R}_q$. $P(k_i|\mathbf{R}_q)$ é a probabilidade do termo k_i estar presente em um documento escolhido aleatoriamente do conjunto $\mathbf{R}_q$. $P(k_i|\overline{\mathbf{R}_q})$ é a probabilidade do termo k_i estar presente em um documento escolhido aleatoriamente do conjunto $\overline{\mathbf{R}_q}$. Seja $P(\mathbf{R}_q|d_j)$ a probabilidade do documento d_j ser relevante para a consulta q e $P(\overline{\mathbf{R}_q}|d_j)$ a probabilidade do documento d_j não ser relevante para a consulta q. No modelo probabilístico, a similaridade entre um documento d_j e uma consulta q é definida por:

$$sim(d_j, q) = \frac{P(\mathbf{R}_q|d_j)}{P(\overline{\mathbf{R}_q}|d_j)} \tag{2.4}$$

Para simplificar o cálculo, o modelo probabilístico assume independência dos termos. Além disso, para calcular as probabilidades $P(\mathbf{R}_q|d_j)$ e $P(\overline{\mathbf{R}_q}|d_j)$, o modelo aplica a elas uma série de transformações preservando a ordem com o objetivo de obter uma estimativa numérica para o *ranking* do documento d_j. Tais transformações de preservação de ordem incluem a aplicação das regras de

Bayes e logaritmos. Como resultado destas transformações, temos:

$$sim(d_j, q) = \sum_{i=1}^{t} w_{i,j} \cdot w_{i,q} \cdot \sigma_{i/R} \tag{2.5}$$

, onde

$$\sigma_{i/R} = \ln \frac{P(k_i|\mathbf{R}_q)}{1 - P(k_i|\mathbf{R}_q)} + \ln \frac{1 - P(k_i|\overline{\mathbf{R}_q})}{P(k_i|\overline{\mathbf{R}_q})} \tag{2.6}$$

Esta é a expressão clássica para determinar o *ranking* no Modelo Probabilístico. Detalhes da derivação desta equação estão em [65].

As principais desvantagens do modelo probabilístico são o fato de que, para várias aplicações, a distribuição dos termos entre documentos relevantes e irrelevantes não estará disponível, o fato de que o método não leva em conta a freqüência com que os termos ocorrem dentro dos documentos e a adoção da abordagem de independência para os termos. A principal vantagem deste modelo, é o fato dos documentos serem ordenados em ordem decrescente de acordo com a probabilidade de serem relevantes [1].

2.2 Redes Bayesianas

Redes Bayesianas [45] (também conhecida como redes de inferência ou redes de crença) produzem bons resultados quando aplicadas à problemas de RI, tanto para simulação de modelos tradicionais de RI, quanto para combinação de informação de diferentes fontes [6, 47]. Estas redes permitem uma visão uniforme, flexível e formal de muitos problemas para combinar fontes de informação ou evidências. Utilizamos a modelagem bayesiana para combinar fontes de evidências usando a estrutura funcional. Iniciamos descrevendo alguns conceitos relacionados à Teoria da Probabilidade. Estes conceitos são fundamentais para definição de Redes Bayesianas. Após apresentar a definição formal de Redes Bayesianas, introduzimos o modelo de Redes de Crença para RI e como representar o modelo Vetorial no modelo de Redes de Crença.

2.2.1 Conceitos Básicos

Teoria da Probabilidade

No mundo em que vivemos lidamos constantemente com a incerteza, ou por não possuirmos informações completas sobre os fatos que nos cercam, ou mesmo por desconhecermos alguns deles. Consciente ou inconscientemente, tomamos decisões com graus de crença baseados em fatos passados ou em regras gerais. Quando dizemos, por exemplo, que a probabilidade de acontecer tal fato é de 90%, estamos exprimindo um grau de crença ou expectativa que tal fato irá acontecer. As duas correntes mais importantes na área de Probabilidades são as correntes freqüentista e epistemológica. A corrente freqüentista defende a posição de que números que representam as probabilidades são provenientes de experimentos. A corrente epistemológica interpreta os números como graus de crença que podem ser obtidos sem experimentação. O mecanismo de inferência na rede bayesiana aplicada em RI é baseado nas regras da teoria da probabilidade e em uma visão epistemológica.

Na teoria da probabilidade temos a probabilidade *a priori* ou incondicional e a probabilidade *a posteriori* ou condicional [50].

A probabilidade *a priori* ou incondicional ocorre quando não conhecemos nenhuma evidência. É denotada por $P(A)$, ou seja, a probabilidade *a priori* da proposição A ser verdadeira. Exemplo: P(Fumante) = 0,1. Significa que, sem conhecer nenhuma informação *a priori*, a probabilidade de uma pessoa ser fumante é de 0,1 ou 10%.

A probabilidade *a posteriori* ou condicional ocorre quando conhecemos as evidências e pode ser definida em termos da probabilidade *a priori*, $P(A|B)$ lê-se probabilidade de A dado que tudo que conhecemos é B, e é dado pela equação:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \tag{2.7}$$

Exemplo: $P(CancerPulmao|Fumante) = 0,6$. Indica que se é observado que um paciente é fumante e não há nenhuma outra informação disponível, então a probabilidade do paciente ter câncer de pulmão é de 0,6 (60% dos doentes analisados até ao momento que são fumantes têm câncer de pulmão).

Axiomas da Probabilidade

1. Todas as probabilidades estão entre 0 e 1, ou seja, $0 \leq P(A) \leq 1$;
2. Proposições necessariamente verdadeiras têm probabilidade 1 e proposições necessariamente falsas têm probabilidade 0, ou seja, $P(\textit{verdadeiro}) = 1$ e $P(\textit{falso}) = 0$;
3. A probabilidade da disjunção entre dois eventos A e B é dada por: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$.

Desses três axiomas podemos derivar todas as propriedades de probabilidade [50].

Distribuição Conjunta de Probabilidade

Um modelo probabilístico de um domínio consiste de um conjunto de variáveis aleatórias que podem ter valores particulares com certas probabilidades. A distribuição conjunta de probabilidade especifica completamente todas as proposições do domínio. Um evento atômico é uma especificação completa do estado do domínio, ou seja, uma atribuição de valores particulares para todas as variáveis. Sejam as variáveis aleatórias $X_1, X_2, \ldots, X_m$. A distribuição conjunta de probabilidade $P(X_1, X_2, \ldots, X_m)$ atribui probabilidades para todos os possíveis eventos atômicos. A distribuição conjunta de probabilidade é uma tabela m-dimensional na qual cada célula fornece a probabilidade de que tal estado específico ocorra. $P(X_i)$ é um vetor uni-dimensional de probabilidades para todos os possíveis valores da variável X_i.

Teorema de Bayes

Pelas duas formas da regra do produto, esquematizada na Figura 2.1, temos:

$$P(A \wedge B) = P(A|B)P(B) \tag{2.8}$$

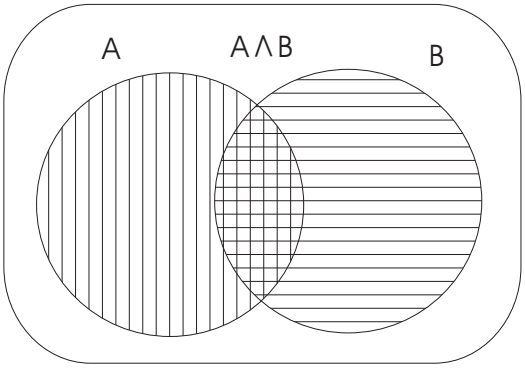

Fig. 2.1: Conjuntos para Representação da Regra do Produto da Teoria da Probabilidade

$$P(A \wedge B) = P(B|A)P(A) \tag{2.9}$$

Podemos escrever:

$$P(A|B)P(B) = P(B|A)P(A) \tag{2.10}$$

Então:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.11}$$

Esta equação é conhecida como regra de $Bayes$ (também lei ou teorema de $Bayes$). Dada uma evidência E, podemos reescrevê-la:

$$P(B|A, E) = \frac{P(A|B, E)P(B|E)}{P(A|E)} \tag{2.12}$$

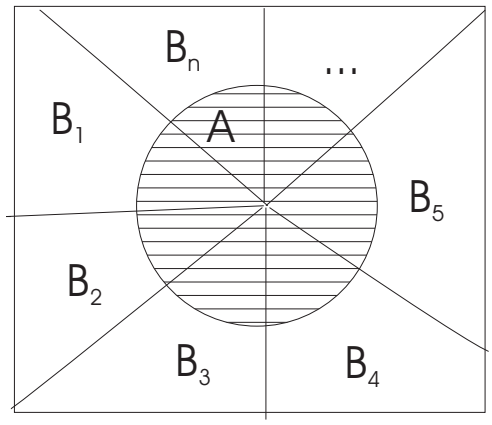

Fig. 2.2: Interpretação gráfica para a Regra da Probabilidade Total

Regra da Probabilidade Total

A regra da probabilidade total diz que qualquer evento A pode ser escrito como a união dos eventos B_i, se $B_i, i = 1, 2, \ldots, n$, é um conjunto de proposições mutuamente exclusivas (chamada de partição) ilustrada na Figura 2.2. Assim, temos:

$$P(A) = P(A \wedge B_1) + P(A \wedge B_2) + \cdots + P(A \wedge B_n) \tag{2.13}$$

$$P(A) = \sum_{i}^{n} P(A \wedge B_i) \tag{2.14}$$

Usando a regra do produto, podemos escrever:

$$P(A) = \sum_{i}^{n} P(A|B_i)P(B_i) \tag{2.15}$$

Dada uma evidência E, podemos reescrevê-la:

$$P(A|E) = \sum_{i}^{n} P(A|B_i, E)P(B_i|E) \tag{2.16}$$

Regra de Independência

Se os eventos A e B são independentes, então temos duas equações:

$$P(A|B) = P(A) \tag{2.17}$$

$$P(A \wedge B|C) = P(A|C)P(B|C) \tag{2.18}$$

Se os eventos A e B são condicionalmente independentes dado C, então temos:

$$P(A|B, C) = P(A|C) \tag{2.19}$$

Redes Bayesianas

As Redes Bayesianas consistem em um grafo acíclico direcionado de dependências, cujos nós representam variáveis randômicas proposicionais ou constantes, e as arestas indicam as relações de dependência entre os nós [45]. No grafo, uma aresta significa que o primeiro tem influência direta sobre o segundo. Esta influência é quantificada através de uma função de distribuição de probabilidade condicional correlacionando os estados de cada nó com os estados dos nós pais. A idéia principal é que, para descrever um modelo do mundo real, não é necessário usar uma enorme tabela de probabilidade conjunta na qual são listadas as probabilidades de todas as combinações possíveis de eventos.

A topologia da rede pode ser vista como uma base de conhecimento abstrata, representando a estrutura dos processos causais no domínio. Uma vez que a topologia da rede está definida, é necessário especificar as probabilidades condicionais para os nós que participam diretamente das relações de dependência. Cada nó possui uma tabela de probabilidade condicional que quantifica a influência que os nós pais têm sobre cada nó filho.

O princípio fundamental é que as dependências conhecidas entre as variáveis aleatórias do domínio são declaradas explicitamente na rede e que a distribuição conjunta de probabilidade pode ser inferida a partir dessas dependências. Os relacionamentos entre os nós, indicados pelos arcos direcionados do grafo, representam dependências causais ou as influências diretas entre as variáveis do domínio. A intensidade dessas influências ou dependências é expressa por probabilidades condicionais associadas

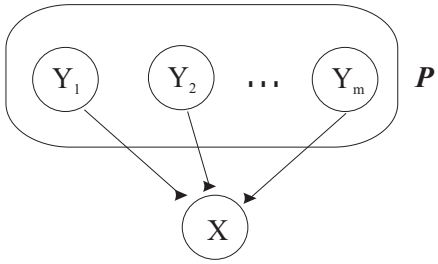

Fig. 2.3: Nós pais de um nó em uma Rede Bayesiana

aos arcos do grafo. As dependências declaradas são utilizadas para inferir as crenças (probabilidades) associadas a todas as variáveis da rede.

Sejam X e Y duas variáveis randômicas e x e y seus respectivos valores. Usamos X e Y para referenciar as variáveis randômicas e os nós na rede associados às variáveis. Um arco direcionado de Y, o nó pai, para X, o nó filho, representa a influência da variável Y sobre a variável X, que é quantificada pela probabilidade condicional $P(x|y)$.

Seja ***P*** o conjunto de nós pais de um nó X, como mostra a Figura 2.3. Seja **p** um conjunto de valores para todas as variáveis em ***P*** e seja x um valor da variável X. A influência de ***P*** sobre X pode ser modelada por qualquer função F tal que $\sum_x F(x, \mathbf{p}) = 1$ e $0 \leq F(x, \mathbf{p}) \leq 1$. A função $F(x, \mathbf{p})$ fornece uma quantificação numérica para $P(x|\mathbf{p})$

Uma rede bayesiana fornece uma completa descrição sobre o seu domínio. Cada entrada na distribuição de probabilidade conjunta pode ser calculada da informação na rede, denotamos por $P(x_1, \ldots, x_r)$, onde r é o número total de variáveis. O valor da entrada é dado pela fórmula $P(x_1, \ldots, x_r)$ = $\prod_{i=1}^{r} P(x_i|Pais(X_i))$.

Um exemplo de uma rede bayesiana de distribuição de probabilidade conjunta $P(x_1, x_2, x_3, x_4, x_5)$ é mostrado na Figura 2.4. O nó X_1, nó raiz, é um nó sem pais cuja distribuição de probabilidade é a probabilidade *a priori* $P(x_1)$. Dado o valor da variável X_1, as variáveis X_2 e X_3 são independentes. Dados os valores das variáveis X_2 e X_3, as variáveis X_4 e X_5 são independentes. Devido à independência declarada na Figura 2.4, a distribuição de probabilidade conjunta pode ser calculada como $P(x_1, x_2, x_3, x_4, x_5) = P(x_1) \cdot P(x_2|x_1) \cdot P(x_3|x_1) \cdot P(x_4|x_2, x3) \cdot P(x_5|x_3)$

Na rede bayesiana, os nós recebem parâmetros numéricos. Estes parâmetros contêm graus de crença de acordo com algum conhecimento. Eles são combinados e manipulados de acordo com os três axiomas básicos da teoria da probabilidade citados anteriormente.

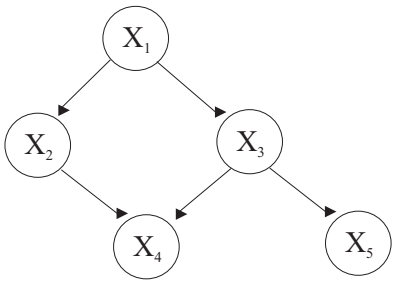

Fig. 2.4: Exemplo de uma Rede Bayesiana

Uma vantagem das redes bayesianas é o poder de síntese de representação dos relacionamentos probabilísticos. É necessário considerar somente o conhecimento de independência entre as variáveis em um domínio. As independências declaradas no tempo de modelagem são usada para inferir crenças para todas as variáveis na rede. O mecanismo de inferência é exponencial em alguns casos, mas é eficiente em muitas situações práticas, particularmente para o contexto de RI. Outra grande vantagem das redes bayesianas é que elas podem ser naturalmente estendidas por evidências geradas a partir de fontes independentes de conhecimento.

2.2.2 O Modelo de Redes de Crença para RI

Nesta seção, descrevemos o Modelo de Redes de Crença proposto em [47] que fornece uma visão epistemológica do problema de RI e interpreta probabilidades como graus de crença.

Redes bayesianas permitem combinar características de diferentes modelos em um mesmo esquema representacional, e por isso, conseguem modelar os eventos e a interdependência de três componentes básicos em RI: palavras-chaves (ou termos), documentos e consultas. Em um modelo probabilístico, cada um desses componentes pode ser visto como um evento. Esses eventos não são independentes, visto que, por exemplo, a ocorrência de um termo influenciará na ocorrência de um documento. Usando uma rede bayesiana, podemos modelar esses eventos e suas interdependências.

Em um SRI tradicional, documentos são indexados por termos. O conjunto de todos os termos é interpretado como um universo K. Seja t o número de termos em uma coleção, então $K = \{k_1, k_2, \ldots, k_t\}$.

Cada termo k_i está associado à uma variável randômica, denotada por K_i. Esta variável é 1 para indicar que um evento associado com o termo k_i ocorreu. Para simplicar a notação, escrevemos $P(k_i)$

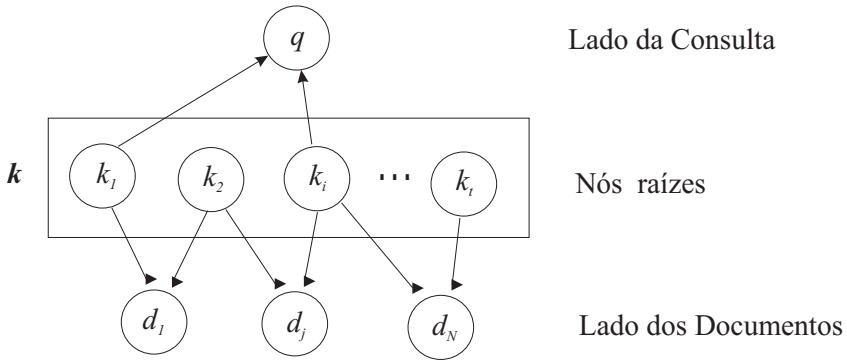

Fig. 2.5: Rede Bayesiana para uma consulta q composta pelos termos k_1 e k_i

ao invés de $P(K_i = 1)$ e $P(\bar{k_i})$ ao invés de $P(K_i = 0)$.

Um documento d_j é modelado como um conjunto composto de termos selecionados que ocorrem em seu texto. Uma variável randômica D_j é associada com cada documento d_j, e uma variável Q com a consulta q do usuário. Considerando também que as consultas são compostas de termos. A rede resultante desta modelagem é mostrada na Figura 2.5.

Nesta rede, cada nó d_j modela um documento, o nó q modela a consulta do usuário, e o nó k_i modela os termos na coleção. A instanciação dos nós raízes separa os nós documentos do nó da consulta, tornando-os mutuamente independente. Então, na rede de crença da Figura 2.5, dizemos que a consulta está do lado da consulta da rede, enquanto os documentos estão do lado dos documentos da rede.

A similaridade entre um documento d_j e a consulta q pode ser interpretada como a probabilidade do documento d_j ser observado dado que a consulta q foi observada. Então, usando a lei de $Bayes$ e a regra da probabilidade total, calculamos a similaridade $P(d_j|q)$ como:

$$P(d_j|q) = \eta \sum_{\forall \mathbf{k}} P(d_j|\mathbf{k}) \times P(q|\mathbf{k}) \times P(\mathbf{k}) \tag{2.20}$$

onde $\eta = 1/P(q)$, como usada em [45], é uma constante de normalização. Esta equação é a expressão genérica para o *ranking* de um documento d_j em relação a consulta q, em um modelo de rede de crença.

Para representar qualquer um dos modelos tradicionais de RI, usando a rede da Figura 2.5,

precisamos apenas definir as probabilidades $P(d_j|\mathbf{k})$,$P(q|\mathbf{k})$ e $P(\mathbf{k})$ apropriada. Como exemplo, mostramos como representar o modelo vetorial, descrito na Seção 2.2.3.

2.2.3 Rede de Crença para o Modelo Clássico Vetorial

O modelo vetorial, muito popular na área de RI, foi introduzido na Seção 2.1.2. Uma rede de crença pode ser usada para calcular um *ranking* do modelo vetorial fazendo a Equação 2.20 equivalente à Equação 2.1. Isto é realizado através da especificação das probabilidades $P(d_j|\mathbf{k})$,$P(q|\mathbf{k})$ e $P(\mathbf{k})$ como:

$$P(\mathbf{k}) = \frac{1}{2^t} \tag{2.21}$$

$$P(q|\mathbf{k}) = \begin{cases} 1, & \text{se } \forall i, K_i = 1, \text{ se e somente se, } k_i \text{ está na consulta } q \\ 0, & \text{caso contrário} \end{cases} \tag{2.22}$$

$$P(d_j|\mathbf{k}) = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,k}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,k}^2}} \tag{2.23}$$

Na Equação 2.21, t é o número de termos da coleção. Podemos definir para esta equação a probabilidade *a priori* $P(\mathbf{k})$ como constante para todo **k**. A Equação 2.22 restringe a computação do estado **k** onde somente os termos da consulta são observados. A Equação 2.23 é definida como a similaridade do cosseno, onde $w_{i,j}$ e $w_{i,k}$ são os pesos usados no modelo vetorial e t é o número total de termos distintos na coleção.

Aplicando as Equações 2.21,2.22 e 2.23 na Equação 2.20, temos:

$$P(d_j|q) = \alpha \times \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2.24}$$

onde α é uma constante que combina η e $P(\mathbf{k})$. Então, o *ranking* de documentos definido pela Equação 2.24 coincide com o *ranking* de documentos definido pelo modelo clássico vetorial.

De modo similar, o modelo de redes bayesianas pode ser usado para construir modelos que combinam informações de *links* e outras evidências. Na próxima seção descrevemos alguns conceitos da análise de *links*.

2.3 Fontes de Evidências: Análise de *Links*

Nesta seção apresentamos alguns conceitos sobre análise de *links* ou ligações.

A estrutura de *links* é uma das fontes de informação ou evidência mais rica na Web [37]. A análise da estrutura de *links* é utilizada pelos mecanismos de busca por constituir uma forma de avaliação humana das páginas. As páginas apontadas por um maior número de *links* tendem a ser mais relevantes do que aquelas pouco referenciadas [8]. A presença de *links* pode determinar o que os autores tentam classificar, unir ou indicar.

Os *links* são evidências que podem ser combinadas e adicionadas aos modelos de RI para tentar melhorar a qualidade do *ranking* [37]. A estrutura de *links* e o texto do *link* provêem informações valiosas tanto para avaliação de relevância como para filtragem da qualidade.

Vários algoritmos foram propostos para extrair informações da estrutura de *links* da Web. De forma geral, estes algoritmos tratam a Web como um grafo direcionado, onde cada nó corresponde a um documento da Web, e cada aresta corresponde ao *link* entre documentos. Os *links* entre os documentos podem derivar informações sobre a importância de determinado documento em relação a um dado assunto.

Alguns desses algoritmos são: HITS [37], PageRank [44] e Salsa [41]. Os algoritmos de análise de *links* analisam a estrutura de *links* de uma coleção de documentos Web com o objetivo de extrair desta coleção, informações que podem ser utilizadas para vários propósitos [8]. Os objetivos dos algoritmos de análise de *links* são:

- associar um peso de "importância" às páginas ou documentos na Web. Os algoritmos que associam pesos de importância aos documentos são chamados de *topic distillation*. A utilização deste peso possibilita aos sistemas de busca retornarem as páginas que além de relevantes à

consulta são também consideradas importantes para o tópico pesquisado. Estes algoritmos podem ser divididos em duas famílias: algoritmos dependentes da consulta (análise de *link* local) e algoritmos independentes da consulta (análise de *link* global).

- identificar comunidades na Web - as comunidades são definidas por um conjunto de páginas que se auto-referenciam e abordam um tópico específico;

- encontrar as páginas similares a uma página Web - funcionalidade "páginas parecidas" encontrada em alguns sistemas de busca ;

- identificar a reputação de uma página na Web;

- elaborar políticas de *crawling* [16]: coleta de páginas na Web;

- classificar documentos Web;

- associar contexto às consultas realizadas submetidas aos sistemas de busca.

Um algoritmo muito utilizado é o algoritmo HITS que utiliza informação proveniente da estrutura de *links* entre os documentos para medir a importância de um determinado documento, de acordo com duas métricas: autoridade e *hub*. Páginas autoridades são páginas que são referenciadas por páginas *hub* e páginas *hub* são páginas que apontam para múltiplas páginas autoridades. O valor de autoridade de determinado documento é definido recursivamente em função do número de apontadores provenientes de documentos com determinados valores de *hub* definidos. A mesma definição é válida para o valor de *hub* de um documento. O valor de *hub* de determinado documento é definido recursivamente em função do número de apontadores para documentos com valores de autoridade definidos.

Os conceitos de *Hub* e Autoridade são recursivos. Um bom *hub* é uma página que aponta para várias boas autoridades. Uma boa autoridade é uma página apontada por vários bons *hubs*. Os valores de autoridade e *hub* são considerados valores de evidências. Veja o exemplo de uma estrutura de *links* na Figura 2.6.

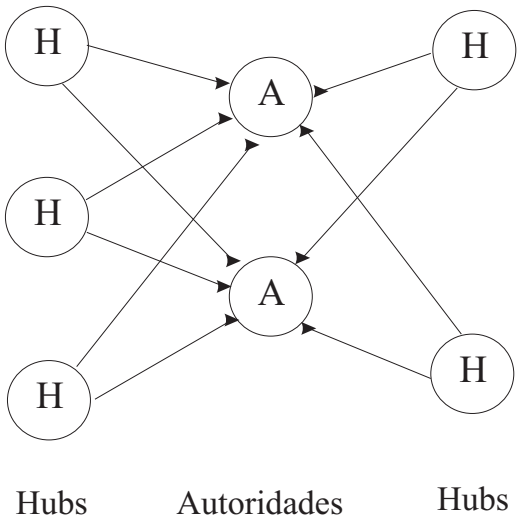

Fig. 2.6: Conjunto de *Hubs* e Autoridades

2.4 Meta-Modelos Alternativas

Com o crescimento e as diferenças entre as estratégias de recuperação, uma grande variedade de meta-modelos formais têm sido propostas. Existem meta-modelos muito genéricos que são utilizados somente em contextos de alta abstração de tarefas de RI. Classificamos os meta-modelos formais em: meta-modelos algébricos, meta-modelos baseados em probabilidades e meta-modelos baseados em lógicas. O modelo de redes de crença apresentado na Seção 2.2 é um tipo de meta-modelo baseado em probabilidades. Alguns meta-modelos algébricos e lógicos são apresentadas nesta Seção.

2.4.1 Meta-Modelos Algébricos

Apresentamos aqui alguns meta-modelos algébricos. O meta-modelo funcional também é classificado como meta-modelo algébrico. Os meta-modelos algébricos apresentados a seguir não apresentam uma forma para realizar a comparação entre modelos de RI como mostramos em nosso trabalho e também não apresentam aplicações de seus meta-modelos para representar outros modelos de RI diferentes dos modelos clássicos.

Meta-Modelo de Grossman e Frieder

Um meta-modelo algébrico muito genérico foi apresentada em [27]. Esse meta-modelo define um modelo de RI com uma tupla:

$$I = \langle D, Q, \delta \rangle$$

onde

- D = conjunto de documentos
- Q = conjunto de consultas
- δ = função de recuperação

A função de recuperação δ produz um subconjunto de documentos δ_i como resposta a uma consulta $q_i \in Q$.

Este meta-modelo é simples e claro, mas a principal desvantagem é que existe uma dificuldade ao usá-lo para representar os modelos clássicos de recuperação (vetorial, probabilístico e booleano), pois não define *ranking* e é muito geral. Este meta-modelo é usado para introduzir conceitos de recuperação, mas tem pouca utilização na prática.

Meta-Modelo Caracterização BR-Formal

Outro meta-modelo algébrico é apresentado em [1] que chamamos de Caracterização BR-Formal. Este meta-modelo é mais rico que a apresentada anteriormente, pois define a função de *ranking*. Para este meta-modelo, um modelo de RI é uma quadrupla:

$$M = \langle D, Q, F, R \rangle$$

onde

- D = conjunto dos documentos
- Q = conjunto de consultas
- F = *framework* para modelagem dos documentos, consultas e seus relacionamentos

- R = é uma função de *ranking*. R: $Q \times D \rightarrow \mathbb{R}$.

A flexibilidade desta caracterização consiste no componente *framework*. Este componente pode ser o espaço vetorial com seus operadores, o conjunto algébrico para o modelo booleano, ou qualquer outro *framework* usado para modelar uma estratégia de recuperação. Este meta-modelo contém todos os componentes de um modelo de RI, mas também muito geral na prática. Ele é tão geral que os autores raramente o utilizam [46]. Além disso, este trabalho não possui aplicações do meta-modelo proposto.

Meta-Modelos de Dominich

Dominich tem um extenso trabalho na formalização de modelos de RI. Os sistemas de RI são estudados do ponto de vista matemático. Ele propõe um *framework* algébrico em [18, 19] para qualquer modelo clássico de RI (Vetorial, Probabilístico e Booleano). O autor introduz alguns conceitos:

Identificadores: parte de informação usada para descrever um documento (termos, palavras chaves, descritores);

Objetos: qualquer parte de informação para compor um documento(texto, imagens,sons,...), pode ser o próprio documento;

Documentos: conjunto de objetos;

Critérios: reflete um relacionamento entre dois documentos (similaridade, relevância, distância);

Limiar: usado quando define o modelo de recuperação. O conjunto de documentos deve satisfazer um critério maior que o limiar;

Recuperação: a recuperação é um mapeamento de um documento para um conjunto de documentos;

Uma recuperação de informação clássica (CIR) é definida como um sistema composto por uma coleção de documentos e um mapeamento chamado recuperação, pela tupla:

$$\langle D, \mathfrak{R} \rangle$$

Onde $\mathfrak{R}$ é uma função que retorna o conjunto resposta. Define duas propriedades: a primeira propriedade é a reflexividade, no caso em que o documento é igual à consulta, então qualquer critério

retorna o valor 1. A segunda propriedade é chamada de critério arbitrário, onde a recuperação será a interseção entre dois conjuntos: um conjunto com os documentos que tem uma similaridade para um dado critério sempre maior que a similaridade retornada por qualquer outro critério e outro para os documentos com um conjunto de similaridades do critério ajustado com um *threshold* (α_i). Este *framework* representa os modelos clássicos vetorial e probabilístico.

Em outro trabalho de *Dominich* [20], linguagens e gramáticas formais são aplicadas para definir outro modelo de recuperação de informação. Ele modela a necessidade de informação do usuário como:

$$\text{IR} = m[\mathbb{R}(O, (Q, \langle I, \vdash \rangle))]$$

onde

- O = conjunto de objetos a serem recuperados
- Q = conjunto de consultas
- I = informação do usuário
- $\vdash$ = informação deduzida da informação do usuário
- $\mathbb{R}$ = relacionamento entre objetos e a necessidade de informação
- m = representa que a relação $\mathbb{R}$ possui alguma incerteza

Este *framework* representa perfis de usuários, desde que a informação do usuário personalizado seja armazenada em ordem para inferir informação adicional para quando especificada sua necessidade. Uma gramática é utilizada para representar documentos e consultas na forma normal disjuntiva, ambos podem ser representados como expressões booleanas por termos e operadores lógicos ($\wedge, \vee, \neg$). Este trabalho não está relacionado diretamente aos meta-modelos apresentados nesta Seção, pois representa apenas perfis de usuários.

Meta-Modelo de Atribuição de Termos

Montejo [46] propõe uma representação formal para sistemas de recuperação de informação. Este meta-modelo é similar aos outros, mas enfatiza a função de *ranking* entre documentos e consultas. Um SRI é modelado por:

$$SRI = \langle D, Q, T, r \rangle$$

onde

- D = conjunto de documentos na coleção
- Q = conjunto de consultas
- T = conjunto de termos das consultas e documentos
- r = função de *ranking* com duas propriedades: reflexividade e simetria

A função de *ranking* é uma função de mapeamento onde a imagem é qualquer par ($texto, texto$), sendo $texto$ qualquer documento ou consulta. Esta função retorna n documentos com os mais altos valores de *ranking*. Além dessas definições, este meta-modelo define o conceito de atribuição de termos, onde uma função de atribuição (ρ), dado um documento produz um conjunto de termos (vocabulário), é definida. Este trabalho apresenta a representação dos modelos clássicos de RI, mas muito simplificada, pois apenas denomina a função de similaridade como função de *ranking*.

Meta-Modelo Baseado em Matriz

Em [63] é apresentado um *framework* geral para modelagem de RI, onde coleções, documentos e consultas correspondem a matrizes no espaço. Aspectos de recuperação, tais como conteúdo, estrutura e semântica, são expressos por matrizes definidas nos espaços de coleções, de documentos e de consultas.

A estrutura matemática proposta pode ser usada para expressar as medidas clássicas e alternativas de avaliação envolvendo, por exemplo, a estrutura dos documentos e para explicar e relacionar modelos de RI. A maior motivação para este *framework* inclui a generalização do modelo vetorial e do modelo probabilístico. O nível de abstração desse meta-modelo é menor que o nível de abstração da estrutura funcional tornando-o mais restrito.

2.4.2 Meta-Modelos Baseados em Lógica

Os meta-modelos baseados em Lógica possuem uma abordagem diferente do meta-modelo funcional. O meta-modelo funcional possui o mesmo nível de expressividade que os metas-modelos baseados

em Lógica. Isso porque é baseada em funções e foi construída visando tornar os modelos de RI mais práticos no sentido de implementação.

Os meta-modelos baseados em Lógica tentam formalizar o estudo de propriedades e as características de modelos RI em um ambiente lógico. Estes meta-modelos são um *framework* uniforme com várias características de sistemas de RI [43].

Meta-Modelos Baseados em *Aboutness*

O uso da Lógica para definir provas para RI foi proposto inicialmente em [43], mostrando que um modelo lógico é uma forma geral de muitos outros modelos de RI. A idéia foi posteriormente investigada em [4], onde um *framework* foi proposto em que diferentes modelos de RI foram teoricamente representados e formalmente estudados. O *framework* foi desenvolvido com uma lógica, permitindo que provas formais sejam realizadas.

O *framework* define a relação *aboutness*, denotado por $\models$, com o objetivo de capturar a informação sobre outra informação para RI, isto é, a informação de relevância. Dado dois objetos a e b, $a \models b$ significa que a é relevante para b. Axiomas são definidos para possivelmente representar as propriedades dos sistemas de RI. Exemplos dos axiomas incluem:

- Reflexividade: $a \models a$
- Simetria: se $a \models b$, então $b \models a$
- Transitividade: Se $a \models b$ e $b \models c$, então $a \models c$

Outro meta-modelo lógico é proposto em [30]. Este trabalho propõe um *framework* para comparação teórica entre modelos de RI baseado na Teoria da Situação. A Teoria da Situação é estuda a relevância entre documentos ou entre um documento e uma consulta. Os modelos são comparados de acordo com alguns axiomas e usando também a propriedade denominada *aboutness* que cada um dos modelos possui. Como pesquisa futura, [30] propõe a definição de um conjunto completo de axiomas e realizar uma prova de completude.

No próximo capítulo discutimos o relacionamento de alguns meta-modelos com o meta-modelo funcional apresentando as vantagens da estrutura funcional.

Capítulo 3

Trabalhos Relacionados

Neste capítulo discutimos os principais trabalhos relacionados à nossa proposta. Na Seção 3.1 apresentamos alguns meta-modelos para RI. Na Seção 3.2 discutimos os trabalhos relacionados à combinação de informações de evidências na Web.

3.1 Meta-Modelos Formais

Os meta-modelos formais na literatura podem ser classificadas como lógicos ou algébricos [46]. Aqui, classificamos os meta-modelos formais em: meta-modelos algébricos, meta-modelos baseados em probabilidades e meta-modelos baseados em lógica. Os meta-modelos baseados em probabilidades são um tipo de meta-modelo algébrico. O meta-modelo proposto neste trabalho pode ser classificada como algébrico.

3.1.1 Meta-Modelos Algébricas

Existem muitos trabalhos sobre meta-modelos formais para RI baseados em considerações algébricas. Uma caracterização formal dos modelos de RI é apresentado em [1]. Nesta caracterização, chamada aqui por Caracterização BR-Formal, são definidos quatro componentes que um modelo deve ter: visão lógica de documentos, visão lógica de consultas, um *framework* para modelagem de documentos, consultas e seus relacionamentos, e uma função de *ranking*. Este modelo é completo e rico, mas é um modelo muito geral e por isso não é utilizado na prática. Já o meta-modelo proposto neste

trabalho, funcional, também define os componentes que um modelo de RI deve possuir e é uma ferramenta para comparar equivalência entre modelos. Também é rica e completa, mas com um nível de abstração menor que o meta-modelo apresentado em [1]. Por exemplo, nosso meta-modelo define as propriedades para a função de similaridade ou *ranking*.

O artigo [46] faz uma revisão de alguns métodos formais para sistemas de RI e propõe um novo meta-modelo formal generalizando as definições de documentos, consultas, função de *ranking* e conjunto recuperação. Este meta-modelo permite a representação dos modelos clássicos. Em nosso meta-modelo, as definições de documentos, consultas, função de *ranking* (similaridade), e conjunto recuperação são definidos, além de outras definições. Além disso, todos os modelos de RI expressos por um algoritmo podem ser representados no meta-modelo funcional. Outra vantagem de nosso trabalho define uma formalização para comparação de equivalência entre modelos.

Outros meta-modelos formais são propostos por Dominich em [18, 19, 20]. Este autor tem realizado um extenso trabalho na formalização de modelos. Em [18, 19], propõe um *framework* definindo alguns conceitos para modelagem de qualquer modelo clássico de RI. Esses trabalhos definem conjunto recuperação, mas não definem *ranking*. O artigo [20] mostra uma definição formal de RI através da medida de uma relação entre documentos e um modelo de usuário, mas não apresenta aplicações práticas para modelos de RI. O trabalho não mostra representação e nem comparação de modelos.

3.1.2 Meta-Modelos Baseados em Probabilidades

Existem alguns trabalhos sobre meta-modelos formais ou *frameworks* genéricos baseados em probabilidades para modelos de RI. Estes *frameworks* são baseados principalmente em redes bayesianas.

As redes bayesianas, introduzida em [45], fornecem um formalismo gráfico para representar independências entre as variáveis de distribuição de probabilidade conjunta.

Turtle e Croft [64] propoem o primeiro modelo de rede bayesiana para RI, onde demonstram que ao estender o modelo básico de rede de inferência com representações booleanas das consultas de usuário poderia se obter um bom desempenho na qualidade do *ranking*.

Um segundo modelo foi proposto por Ribeiro (Ribeiro-Neto) e Muntz [47], denominado modelo de redes de crença para RI, que é derivado de considerações probabilísticas. Nesta proposta, a aplicação do modelo de redes de crença para modelos de RI é realizada. Já o meta-modelo funcional faz

a aplicação de uma linguagem funcional para os modelos de RI.

Os documentos, termos e consultas são representados no modelo de crença por variáveis aleatórias binárias e o cálculo do grau de relevância é baseado em probabilidades. Além disso, para simplificar a modelagem, na rede de crença os termos são considerados independentes entre si. Por outro lado, no meta-modelo funcional, os documentos, termos e consultas são representados por funções, e o cálculo do grau de relevância é dado pela função de similaridade e os termos podem ser modelados de forma independente ou não.

No modelo de redes de crença [47], a consulta e o documento são modelados do mesmo modo para facilidar a definição da estrutura da rede. Já no meta-modelo funcional, a consulta e o documento podem ser modelados de modo diferente.

Em [54] é proposto um modelo que representa os três modelos clássicos de RI (vetorial, booleano e probabilístico), os ciclos de realimentação de relevantes e alternativas de similaridade consulta-consulta. Outros modelos de RI existentes também podem ser representados através do modelo de redes de crença. Neste trabalho, o meta-modelo proposto é capaz de representar além dos modelos clássicos, todos os modelos que podem ser expressos por um algoritmo.

Outra vantagem do meta-modelo funcional em relação às redes de crença é que o conceito de funções na comunidade de Ciência da Computação é mais usado e mais abrangente que o conceito de probabilidades.

3.1.3 Meta-Modelos Baseados em Lógica

Alguns trabalhos usam Lógica para definir um meta-modelo para modelos de RI. O artigo [11] mostra um resumo de como pesquisas passadas têm combinado o uso de Lógica e incertezas para formulação de modelos de RI.

O uso da Lógica em RI fornece a capacidade para formulação de modelos genéricos e torna possível o estudo de propriedades desses modelos. Meta-modelos lógicos para IR são estudados para fornecer uma rica e uniforme representação da informação e sua semântica. Geralmente, em um modelo lógico as consultas e documentos podem ser representados por fórmulas lógicas. A inferência é associada com implicação lógica: um documento é relevante para a consulta significa que o documento implica na consulta. Somente a Lógica não é capaz de representar um modelo de

RI, a teoria da incerteza é necessária [11].

Existem outros estudos com abordagens lógicas sobre meta-modelos em RI. O uso da Lógica para formalmente conduzir provas para RI foi proposto inicialmente em [43]. Na década passada, diversos meta-modelos têm sido propostos [5, 31, 32]. Meta-Modelos lógicos podem ser classificados em três tipos: baseados na teoria da situação, baseados em lógica modal, e outros tipos [39, 59, 70].

Nosso meta-modelo também permite a formulação de modelos genéricos e o estudo das características desses modelos, pórem utiliza outra abordagem: representa os componentes dos modelos e relações relevantes em um ambiente funcional.

3.2 Combinando Informação de Evidências para RI

Pesquisas recentes tratam da recuperação de documentos em máquinas de busca da Web, utilizando novas fontes de evidências para melhorar a qualidade do *ranking*. O problema de combinar diferentes fontes de evidências em sistemas de RI foi explorada inicialmente no trabalho [51] onde a informação de referências cruzadas era usada para combinar citações com palavras-chaves em coleções compostas por publicações científicas. Contudo, desde a criação da WWW, este assunto tem sido amplamente estudado por muitos pesquisadores em trabalhos recentes [6, 22, 54, 69].

Existem trabalhos sobre combinação de partes de evidências baseadas em conteúdo e em *links* em um único modelo de RI. O artigo [54] utiliza redes de crença para representar e combinar informações baseadas em conteúdo e *link*.

[33] estende o modelo proposto em [54] apresentando uma generalização deste modelo para combinar múltiplas fontes de evidências na rede de crença. Este trabalho tem por objetivo estudar o emprego de um modelo de redes bayesianas como uma alternativa para resolver o problema da combinação de múltiplas fontes de evidência para *ranking* de documentos.

Em [55], a combinação de conteúdo e *link* no contexto do modelo vetorial é apresentada. O primeiro passo de algumas definições do meta-modelo funcional foi apresentado em [55]. Este trabalho tem como seu foco o uso de alguns conceitos de redes bayesianas e da estrutura funcional para propor um modelo vetorial estendido com informações de *links*. Entretanto, a estrutura funcional apresentada tem algumas limitações e falhas, por exemplo, a propriedade de desigualdade triangular

da função de similaridade. Dessa forma, o modelo torna-se restrito, pois a função de similaridade do cosseno não possui a propriedade de desigualdade triangular, então este modelo não obtém a representação para modelo clássico vetorial. Um outro problema é na definição da equivalência que considera dois modelos equivalentes se os valores de suas funções de similaridades sejam iguais, tornando assim a comparação restrita. Em nosso trabalho, a estrutura funcional é aperfeiçoada com o objetivo de mostrar a estrutura funcional como um meta-modelo para modelos de RI. Aqui, removemos a propriedade da desigualdade triangular, melhoramos o nível de notação adicionando novas definições tais como objetos funcionais, coleção de referência funcional, função de documentos relevantes, função de recuperação e topo do *ranking*, e removemos a definição de sobreposição, pois não possui aplicações práticas em RI. Além disso, modificamos a definição de equivalência para considerar a equivalência entre dois modelos funcionais se eles sempre produzirem o mesmo *ranking*. Representamos aqui os modelos clássicos no meta-modelo funcional como é mostrado na maioria dos trabalhos sobre meta-modelos e construímos quatro novos modelos de RI: um modelo baseado em distância equivalente ao modelo clássico vetorial; um modelo vetorial para combinar múltiplas evidências utilizando a função de similaridade noisy-OR do contexto de redes bayesianas; um modelo vetorial para combinar múltiplas evidências utilizando a função de similaridade cosseno; e um modelo de redes de crença para combinar múltiplas evidências utilizando a similaridade do cosseno. Em adição, o artigo [55] não apresenta um trabalho relacionado à meta-modelos, nem vantagens e motivações do meta-modelo funcional como apresentamos nesta dissertação.

Apresentamos dois estudos de caso utilizando o meta-modelo funcional para combinar evidências no contexto de redes de crença e no modelo vetorial. No primeiro estudo de caso, usamos o modelo proposto em [33] generalizando o operador usado na composição dos documentos e projetamos para o modelo vetorial correspondente. Este modelo projetado usa a mesma semântica do modelo proposto em [55], mas generaliza para múltiplas evidências. No segundo estudo de caso, usamos o modelo vetorial para combinar múltiplas evidências encontrado no primeiro estudo de caso modificando a função de similaridade para a similaridade do cosseno e então, construimos um modelo de redes de crença correspondente através do uso do meta-modelo funcional.

Neste capítulo apresentamos os trabalhos relacionados à estrutura funcional e à sua aplicação em modelos que combinam múltiplas evidências. Nosso trabalho difere dos outros pelos seguintes

fatores: mostramos uma aplicação do meta-modelo funcional para combinação genérica de partes de evidências em dois contextos diferentes e de duas formas diferentes. Esta combinação pode ser feita de várias formas em qualquer contexto de modelagem. Uma ferramenta para ajudar desenvolvedores na realização dessa tarefa é o meta-modelo funcional, que unifica modelos. Apresentamos o meta-modelo funcional no próximo capítulo. Além disso, discutimos a capacidade do meta-modelo proposto representar, construir, combinar modelos de RI e comparar modelos similares ou não sem realizar experimentação nos dois próximos capítulos.

Capítulo 4

Estrutura Funcional para RI

Propomos neste trabalho um meta-modelo para RI, chamado Estrutura Funcional para RI. Esta estrutura nos permite representar, construir, combinar e comparar modelos de RI de forma algébrica sem realizar experimentos. Nosso modelo tem expressividade suficiente para generalizar modelos de RI. O alto nível de abstração desta estrutura facilita a construção de modelos e combinação entre eles. Formalmente definimos os componentes tais como documentos, consultas, função de *ranking* ou função de similaridade de um modelo de recuperação de informação.

Nosso *framework* permite mostrar a noção de equivalência entre modelos. A passagem dos modelos para o *framework* funcional facilita a comparação entre eles, pois os modelos são representados usando a mesma linguagem. Os componentes são definidos com base em funções. Ele generaliza todos os modelos de RI que podem ser expressos por um algoritmo, pois é baseado em funções. Então, o meta-modelo é caracterizado por sua simplicidade e o formalismo através de funções pode ser usado para combinar e desenvolver novos modelos.

4.1 Fundamentos da Estrutura Funcional

Apresentamos aqui as definições da estrutura funcional. Os fundamentos são divididos em representação e comparação de modelos.

4.1.1 Representação de Modelos

Para representar modelos de RI na estrutura funcional, definimos termo funcional, função peso, objetos funcionais, função similaridade entre dois objetos funcionais, casamento entre documentos e consultas funcionais, função de documentos relevantes, coleção de referência funcional, função de recuperação e topo do *ranking*.

Definição 1. *Termo Funcional. Um termo funcional é uma função cuja semântica relaciona um conjunto de termos. Um termo funcional f é denotado por $f(k_l, ..., k_s)$, onde $k_l, ..., k_s$ são termos.*

Os termos são palavras-chaves cujo conceito é o mesmo dos modelos clássicos de RI. Seja $\mathbf{K} = \{k_1, \ldots, k_t\}$ um conjunto de termos e $2^{\mathbf{K}}$ o conjunto de sub-conjuntos de **K** chamado conjunto potência. Por exemplo, a função $syn : \mathbf{K} \rightarrow 2^{\mathbf{K}}$ é a função sinônima tal que, dado um termo, retorna o conjunto de sinônimos de cada termo. A função $syn(k_i) = \{k_{i1}, \ldots, k_{is}\}$ retorna o conjunto de sinônimos do termo k_i.

Dado **K**, qualquer função cujo domínio é **K** é um termo funcional. Então, um termo funcional é uma função que expressa qualquer relação entre os termos, sendo esta uma importante ferramenta para modelagem de problemas em RI.

A função peso é um exemplo de termo funcional. Uma função peso é uma função cujo resultado é o peso do termo em um documento ou em uma consulta. Seja $\mathbf{C} = \{d_1, \ldots, d_z\}$ uma coleção de documentos, $\mathbf{K} = \{k_1, \ldots, k_t\}$ um conjunto de termos de **C**, e q uma consulta. A função peso $g : \mathbf{K} \times \{\mathbf{C} \cup \{q\}\} \rightarrow \mathbb{R}$ é tal que $g(k_i, d_j)$ retorna o peso associado ao par (k_i, d_j) e $g(k_i, q)$ retorna o peso associado ao par (k_i, q). Para simplificar, usamos a seguinte notação. Seja $g_j : \mathbf{K} \rightarrow \mathbb{R}$ uma função unária que retorna o peso de um termo no documento d_j. $g_j(k_i)$ retorna o peso associado ao par (k_i, d_j). Analogamente, seja $g_q : \mathbf{K} \rightarrow \mathbb{R}$ uma função unária que retorna o peso de um termo na consulta q. $g_q(k_i)$ retorna o peso associado ao par (k_i, q). As funções peso g_j e g_q são termos funcionais.

Definição 2. *Objetos Funcionais. Um documento funcional df_j é representado por um conjunto de termos funcionais que relacionam os termos do documento d_j. Uma consulta funcional qf é representada por um conjunto de termos funcionais que relacionam termos da consulta q. Objetos funcionais são documentos funcionais ($\boldsymbol{D}_f$) e consultas funcionais ($\boldsymbol{Q}_f$), denotado por $\boldsymbol{O}_f$ =*

$\{df_1, \ldots, df_n, qf_1, qf_2, \ldots, qf_m\}$. *Estes objetos são representados por um conjunto de termos funcionais.*

Definição 3. *Função Similaridade entre dois Objetos Funcionais. Dado um conjunto de objetos funcionais* $\boldsymbol{O}_f = \{df_1, \ldots, df_n, qf_1, qf_2, \ldots, qf_m\}$*, a similaridade é uma função* Δ*:* $\boldsymbol{O}_f \times \boldsymbol{O}_f \rightarrow \mathbb{R}$ *tal que* $\Delta(of_j, of_i) \in \mathbb{R}$ *para cada par* (of_j, of_i)*, onde* $\{of_j, of_i\} \subseteq \boldsymbol{O}_f$*, e satisfaz as seguintes propriedades (ou axiomas):*

1. $0 \leq \Delta(of_j, of_i) \leq 1$ *(normalização)*
2. $\Delta(of_j, of_j) = 1$ *(reflexividade)*
3. $\Delta(of_j, of_i) = \Delta(of_i, of_j)$ *(simetria)*

A função similaridade ou função de *ranking* relaciona termos funcionais de objetos funcionais. Note que neste caso a função de similaridade não necessariamente representa uma função de distância ou métrica (a propriedade de desigualdade triangular não é obrigatória). O modelo vetorial clássico, usando a similaridade do cosseno, por exemplo, não satisfaz a propriedade de desigualdade triangular, satisfazendo apenas as propriedades de simetria, reflexividade e normalização.

A propriedade de normalização da função similaridade é importante para a combinação de evidências e combinação entre modelos, e as propriedades de reflexividade e simetria são importantes para clusterização de documentos. Quando a função de similaridade de um modelo de RI não satisfaz uma dessas propriedades, então ao representar um modelo na estrutura funcional, devido ao poder de abstração das funções, podemos realizar algumas modificações na função similaridade para satisfazer as três propriedades.

Definição 4. *Modelo Funcional. Um modelo funcional é definido pela tupla*

$$\Psi = \langle \boldsymbol{D}_f, \boldsymbol{Q}_f, \boldsymbol{T}_f, \Delta \rangle$$

- $\boldsymbol{D}_f$ *= conjunto de documentos funcionais* $\{df_1, \ldots, df_n\}$
- $\boldsymbol{Q}_f$ *= conjunto finito de consultas funcionais* $\{qf_1, \ldots, qf_m\}$
- $\boldsymbol{T}_f$ *= conjunto de termos funcionais dos documentos e das consultas funcionais* $\{g_1, \ldots, g_v\}$

- Δ = *uma função similaridade, que satifaz as três propriedades acima, a saber: normalização, reflexividade e simetria.*

onde n e m são o número de documentos funcionais e o número de consultas funcionais da coleção de referência, respectivamente.

Realizamos duas simplificações na notação do modelo funcional, sem perda do poder de representação, a saber: o uso de um conjunto unitário de uma consulta funcional e a omissão do $\mathbf{T}_f$. Podemos utilizar um conjunto com apenas uma consulta funcional, pois em geral, nos SRIs as consultas são *ad hoc*, onde a pré-computação não pode ser antecipada. Os termos funcionais podem ser omitidos da notação, pois eles são extraídos dos documentos e das consultas funcionais. Então, usaremos a seguinte simplificação: um modelo funcional é representado por uma coleção de documentos funcionais, um conjunto com apenas uma consulta funcional e uma função de similaridade. Isto é denotado por

$$\Psi = \langle \{df_1, \ldots, df_n\}, \{qf\}, \Delta \rangle,$$

onde Δ é uma função de similaridade sob pares de documentos funcionais ou de um documento funcional da coleção e uma consulta funcional.

Definição 5. *Função de documentos relevantes. A função de documentos relevantes ou conjunto ideal é uma função que dado as consultas funcionais e documentos funcionais retorna os documentos relevantes. Seja $2^{\boldsymbol{D}_f}$ o conjunto potência de documentos funcionais. A função de documentos relevantes é definida por $I : \boldsymbol{Q}_f \times \boldsymbol{D}_f \rightarrow 2^{\boldsymbol{D}_f}$.*

Definição 6. *Coleção de Referência Funcional. A coleção de referência funcional, $\boldsymbol{C}_f$, é formada por um conjunto de objetos funcionais (consultas funcionais $\boldsymbol{Q}_f = \{qf_1, qf_2, \ldots, qf_m\}$ e documentos funcionais $\boldsymbol{D}_f = \{df_1, \ldots, df_n\}$) e pela função de documentos relevantes $I : \boldsymbol{Q}_f \times \boldsymbol{D}_f \rightarrow 2^{\boldsymbol{D}_f}$ ou conjunto ideal para as consultas funcionais. $\boldsymbol{C}_f = \langle \boldsymbol{D}_f, \boldsymbol{Q}_f, I \rangle$*

Definição 7. *Função de Recuperação. Seja $2^{\boldsymbol{D}_f}$ o conjunto potência dos documentos funcionais. A função de recuperação retorna a lista de documentos ordenados (ou ranking) de acordo com a*

função de similaridade (Δ) que são relevantes para a consulta. Esta função é definida por $Rank$: $\boldsymbol{Q}_f \times \boldsymbol{D}_f \rightarrow < 2^{\boldsymbol{D}_f} >$.

Definição 8. *Casamento entre Documentos e Consultas Funcionais. A função similaridade Δ define um ranking cuja ordenação é decrescente. Seja df_j um documento funcional e qf uma consulta funcional. Seja α um número positivo, tal que $0 \leq \alpha \leq 1$. Dado um limite inferior α, o casamento entre df_j e qf ocorre se $\Delta(qf, df_j) \geq \alpha$, onde $\Delta(qf, df_j)$ é uma função similaridade de um modelo funcional.*

Definição 9. *Topo do Ranking. A função topo do ranking pode ser definida de duas formas. A primeira utiliza o conceito de casamento entre documentos e consultas funcionais (Definição 8) Dado um limiar α, a função topo retorna todos os documentos onde $\Delta(qf, df_j) \geq \alpha$, $\forall qf, df_j; 1 \leq j \leq N$. Neste caso, a função topo do ranking Top_α é tal que Top_α: $\mathbb{R} \times < 2^{\boldsymbol{D}_f} > \rightarrow < 2^{\boldsymbol{D}_f} >$.*
A segunda forma considera o topo como sendo uma função que retorna o conjunto de n documentos funcionais do topo do ranking com maiores valores da função de similaridade. Dado o número de documentos que deseja retornar e o ranking definido por $< 2^{\boldsymbol{D}_f} >$, então a função retorna o topo do ranking e é definida por Top_n: $\mathbb{N} \times < 2^{\boldsymbol{D}_f} > \rightarrow < 2^{\boldsymbol{D}_f} >$

4.1.2 Comparação de Modelos

Com a representação de modelos de RI na estrutura funcional, podemos verificar a equivalência entre eles. Definimos uma relação de comparação entre modelos: a relação de equivalência indicada a seguir.

Definição 10. *(Equivalência entre Modelos Funcionais em Relação a uma Consulta). Dois modelos funcionais $\Psi_a = \langle \{df_{a1}, ..., df_{an}\}, \{qf_a\}, \Delta_a \rangle$ e $\Psi_b = \langle \{df_{b1}, ..., df_{bn}\}, \{qf_b\}, \Delta_b \rangle$ são equivalentes em relação à uma consulta qf, se e somente se existe uma função bijetora $\phi : \{df_{a1}, ..., df_{an}\} \rightarrow \{df_{b1}, ..., df_{bn}\}$, tal que se $\phi(df_{ai}) = df_{bi}$ e $\phi(df_{ak}) = df_{bk}$ apresentados na Figura 4.1, então as duas condições abaixo são satisfeitas:*

1. $\Delta_a(qf, df_{ai}) = \Delta_a(qf, df_{ak}) \Leftrightarrow \Delta_b(qf, df_{bi}) = \Delta_b(qf, df_{bk})$
2. $\Delta_a(qf, df_{ai}) > \Delta_a(qf, df_{ak}) \Leftrightarrow \Delta_b(qf, df_{bi}) > \Delta_b(qf, df_{bk})$

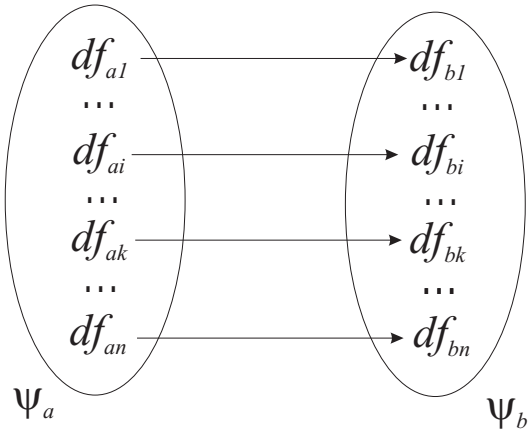

Fig. 4.1: Representação do mapeamento entre os modelos Ψ_a e Ψ_b da definição de equivalência

A propriedade 1 garante que se dois documentos funcionais (df_{ai} e df_{ak}) possuem a mesma similaridade em relação à uma consulta funcional para o modelo Ψ_a, então os mesmos documentos representados no modelo Ψ_b (df_{bi} e df_{bk}) também possuem similaridade iguais em relação à uma consulta funcional para o modelo Ψ_b, ou vice-versa. A propriedade 2 garante que se um documento funcional (df_{ai}) possui similaridade maior que outro documento (df_{ak}) para o modelo Ψ_a, então o primeiro documento (df_{ai}) representado no modelo Ψ_b, por df_{bi}, possui similaridade maior que o segundo documento (df_{ak}) representado no modelo Ψ_b, por df_{bk}. Isto garante que a ordenação do $ranking$ seja a mesma. Com estas duas propriedades, temos que os modelos Ψ_a e Ψ_b geraram o mesmo $ranking$.

Definição 11. *(Equivalência entre Modelos Funcionais). Dois modelos funcionais* $\Psi_a = \langle \{df_{a1}, ..., df_{an}\}, \{qf_a\}, \Delta_a \rangle$ *e* $\Psi_b = \langle \{df_{b1}, ..., df_{bn}\}, \{qf_b\}, \Delta_b \rangle$ *são equivalentes se e somente se* $\forall$ *consulta funcional* $qf \in \boldsymbol{Q}_{fa} \cup \boldsymbol{Q}_{fb}$, Ψ_a *equivale a* Ψ_b *em relação à consulta* qf.

Essas condições garantem que os modelos Ψ_a e Ψ_b sejam equivalentes se e somente se eles geram o mesmo *ranking* quando aplicados a conjuntos de igual tamanho. Neste caso, dois modelos funcionais são equivalentes independente de qualquer consulta funcional qf do conjunto de consultas funcionais de Ψ_a e de Ψ_b.

Representar um modelo na estrutura funcional significa que suas funções de similaridade e forma de representação de documentos e consultas sejam traduzidos na linguagem funcional. O objetivo dessa representação é obter o modelo no formalismo da estrutura funcional. A Figura 4.2 mostra um

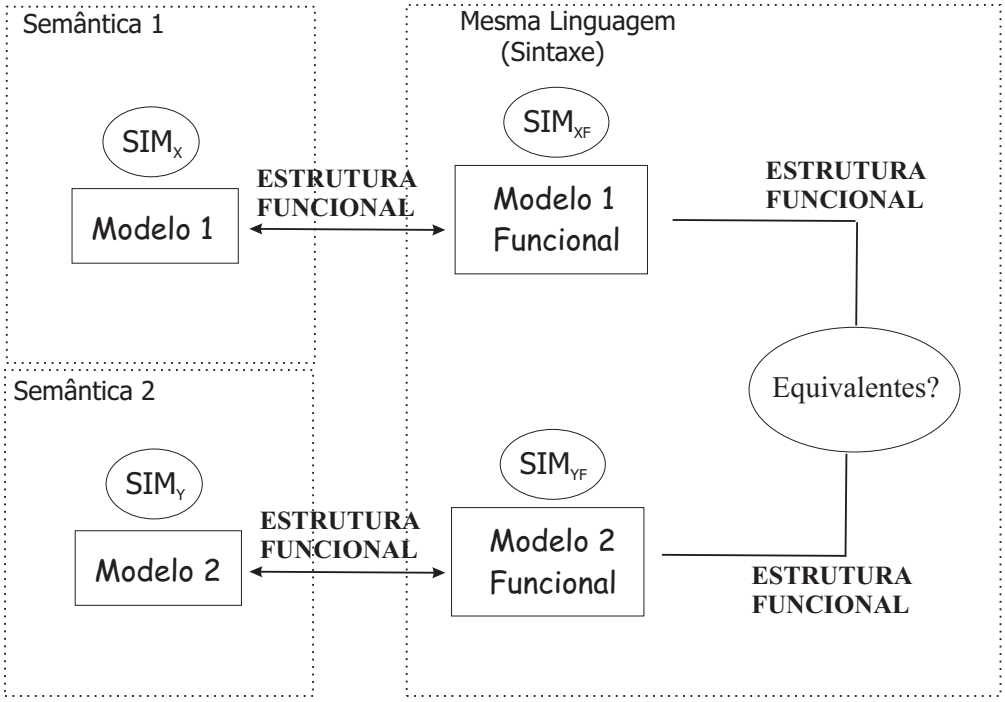

Fig. 4.2: Esquema geral para comparação de equivalência entre modelos de RI

esquema da comparação de equivalência entre os modelos. Dois modelos 1 e 2 contendo semânticas diferentes são traduzidos através da estrutura funcional para os modelos funcionais 1 e 2, respectivamente. Assim, ambos estarão representados em uma mesma linguagem e podemos verificar algebricamente se são equivalentes ou não.

A comparação de equivalência entre os modelos é importante para reutilização de código ou escolha da implementação de um modelo e melhor entendimento da semântica dos modelos. Conforme mostramos na Seção 4.3 existem modelos que mesmo tendo natureza e representação diferentes, quando modelados na estrutura funcional, verifica-se que são equivalentes.

4.2 Representação dos Modelos Clássicos

Nesta seção mostramos como os modelos clássicos [1, 23, 53] são representados na estrutura funcional. Em um modelo clássico, um documento é um registro de dado que inclue uma parte textual. Um termo é uma simples palavra cuja semântica ajuda lembrar o tema principal de um documento.

4.2.1 Modelo Funcional Vetorial

O modelo clássico vetorial foi apresentado na Seção 2.1.2. Aqui representamos suas principais características e o traduzimos para o meta-modelo funcional.

Modelo Clássico Vetorial

No modelo clássico vetorial, consultas e documentos são representados no espaço t-dimensional, onde t é o número de termos da coleção. O conjunto resposta é um *ranking* de documentos construído por uma operação entre o vetor documento e o vetor consulta que define o grau de similaridade entre eles. Uma consulta q e um documento d_j são vetores representados por $\vec{q} =(w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ e $\vec{d_j} = (w_{i,j}, w_{2,j}, \ldots, w_{t,j})$, onde $w_{i,j}$ e $w_{i,q}$ são pesos associados aos termos k_i no documento d_j e na consulta q, respectivamente.

O grau de similaridade entre um documento d_j e uma consulta q no modelo vetorial é uma correlação entre os vetores $\vec{d_j}$ e $\vec{q}$, que pode ser calculado através do cosseno entre esses dois vetores. A *fórmula da similaridade* é definida por:

$$sim_v(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{4.1}$$

O modelo ordena os documentos de acordo com o grau de similaridade para a consulta calculado pela Equação (4.1). Deste modo, os documentos mais similares à consulta ficam no topo do *ranking*.

Modelo Funcional Vetorial

Para representar um modelo na estrutura funcional, é necessário definir um modelo funcional Ψ que o represente. A representação do Modelo Vetorial na estrutura funcional é denotada por Ψ_v, onde $\Psi_v = \langle \{df_{v1}, ..., df_{vn}\}, \{qf_v\}, \Delta_v \rangle$ e,

- $df_{vj} = \{g_j\}$. Os documentos funcionais são conjuntos unários que contêm apenas a função peso para o documento d_j. A função g_j define o peso $w_{i,j}$ no modelo vetorial. Então, $g_j(k_i) = w_{i,j}$;
- $qf_v = \{g_q\}$. As consultas funcionais são conjuntos unários que contêm apenas a função peso da consulta q. No modelo vetorial, a função que define o peso de cada termo na consulta é $w_{i,q}$.

Então, $g_q(k_i) = w_{i,q}$;

- A função similaridade é dada por

$$\Delta_v(df_{vj}, qf_v) = \frac{\sum_{i=1}^{t} g_j(k_i) \cdot g_q(k_i)}{\sqrt{\sum_{i=1}^{t} g_j(k_i)^2} \times \sqrt{\sum_{i=1}^{t} g_q(k_i)^2}} \tag{4.2}$$

Note que esta função similaridade satisfaz as propriedades de normalização, reflexividade e simetria. A propriedade de normalização é válida, pois a função cosseno entre dois vetores cujas coordenadas são positivas retorna um valor entre 0 e 1 ($0 \leq cos(\vec{a}, \vec{b}) \leq 1$). A propriedade reflexividade é válida, pois $cos(\vec{a}, \vec{a}) = 1$ e a propriedade da simetria também é válida, pois $cos(\vec{a}, \vec{b}) = cos(\vec{b}, \vec{a})$.

4.2.2 Modelo Funcional Booleano

O modelo clássico booleano foi apresentado na Seção 2.1.1. Aqui representamos suas principais características e o traduzimos para o meta-modelo funcional.

Modelo Booleano

No Modelo Booleano, documentos são representados por um conjunto de termos e consultas são representados como termos ligados por conectivos OR, AND e/ou NOT. Um peso é associado a cada par termo documento (k_i, d_j) denotado por $w_{i,j}$, onde $w_{i,j} \in \{0, 1\}$, $w_{i,j}$ = 1 se $k_i \in d_j$ e $w_{i,j}$ = 0, caso contrário.

O conjunto resposta gerado é o conjunto de documentos que satisfazem a consulta. Um documento é considerado relevante ou não relevante.

Modelo Funcional Booleano

A representação do Modelo Booleano na estrutura funcional é denotada por Ψ_b, onde $\Psi_b = \langle \{df_{b1}, ..., df_{bn}\}, \{qf_b\}, \Delta_b \rangle$ é um modelo funcional e,

- $df_{bj} = \{g_j\}$. Os documentos funcionais são conjuntos unários que contêm o termo funcional g_j que representa a função booleana da conjunção de: termos pertecentes ao documento d_j e a negação dos termos pertecentes ao documento d_j, isto é, $g_j(k_1, \ldots, k_t) = \bigwedge_{\forall w_{i,j}=1} k_i \bigwedge_{\forall w_{i,j}=0} \neg k_i$.

Suponha por exemplo, os termos do vocabulário a, b, c, d, tais que, $w_{1,j} = 0$, $w_{2,j} = 1$, $w_{3,j} = 0$, $w_{4,j} = 1$. Neste caso, $g_j = \neg a \wedge b \wedge \neg c \wedge d$.

- $qf_b = \{g_q\}$. As consultas funcionais são conjuntos unários que contêm o termo funcional g_q que retorna função booleana dos termos que compõem a consulta clássica q representada na forma normal disjuntiva (FND). Onde $g_q(i)$ ($1 \leq i \leq s$) é uma função que retorna a i-ésima expressão conjuntiva da forma normal disjuntiva de q e s é o número de expressões conjuntivas de FND(q). Isto significa que se a, b, c, d são termos do vocabulário e $q = (\neg a \wedge b \wedge d) \vee c$, então a representação de q na forma normal disjuntiva é dada por
  g_q = $(\neg a \wedge b \wedge d \wedge \neg c) \vee (\neg a \wedge b \wedge d \wedge c) \vee (a \wedge \neg b \wedge d \wedge c) \vee (\neg a \wedge \neg b \wedge d \wedge c) \vee (a \wedge b \wedge d \wedge c) \vee (\neg a \wedge b \wedge \neg d \wedge c) \vee (a \wedge \neg b \wedge \neg d \wedge c) \vee (\neg a \wedge \neg b \wedge \neg d \wedge c) \vee (a \wedge b \wedge \neg d \wedge c)$
  e as expressões conjuntivas dadas por
  $g_q(1) = (\neg a \wedge b \wedge d \wedge \neg c)$, $g_q(2) = (\neg a \wedge b \wedge d \wedge c)$, $g_q(3) = (a \wedge \neg b \wedge d \wedge c)$, $g_q(4) = (\neg a \wedge \neg b \wedge d \wedge c)$, $g_q(5) = (a \wedge b \wedge d \wedge c)$, $g_q(6) = (\neg a \wedge b \wedge \neg d \wedge c)$, $g_q(7) = (a \wedge \neg b \wedge \neg d \wedge c)$, $g_q(8) = (\neg a \wedge \neg b \wedge \neg d \wedge c)$ e $g_q(9) = (a \wedge b \wedge \neg d \wedge c)$.

- Função similaridade é dada por

$$\Delta_b(df_{bj}, qf_b) = \begin{cases} 1, & \text{se } df_{bj} \text{ satisfaz } qf_b \\ 0, & \text{caso contrário} \end{cases} \tag{4.3}$$

Onde df_{bj} satisfaz qf_b se e somente se $\{g_j(k_1, \ldots, k_t)\} \cap \{g_q(1), \ldots, g_q(s)\} \neq \emptyset$. Note que a função similaridade satisfaz as propriedades normalização, reflexividade e simetria. A propriedade de normalização é válida, pois a função de similaridade possui apenas valores 0 ou 1. A propriedade de reflexividade é válida, pois o documento funcional df_{bj} satisfaz ele mesmo, ou seja $\Delta_b(df_{bj}, df_{bj}) = 1$, e a propriedade de simetria é válida, pois $\Delta_b(df_{bj}, qf_b) = \Delta_b(qf_b, df_{bj})$.

4.2.3 Modelo Funcional Probabilístico

O modelo clássico probabilístico foi apresentado na Seção 2.1.3. Aqui representamos suas principais características e o traduzimos para o meta-modelo funcional.

Modelo Probabilístico

Documentos e consultas no Modelo Clássico Probabilístico são representados por um conjunto de termos. Existe um conjunto de documentos relevantes (conjunto ideal). Pesos $w_{i,j} \in \{0,1\}$ e $w_{i,q} \in \{0,1\}$ são associados a cada par (k_i, d_j) e (k_i, q), respectivamente. Seja R uma estimativa para o conjunto ideal e $\neg R$ seu complemento. A função similaridade é definida por

$$sim_p(d_j, q) = P(R|d_j)/P(\neg R|d_j) \tag{4.4}$$

onde, $P(R|d_j)$ é a probabilidade no Modelo Probabilístico do documento ser relevante e $P(\neg R|d_j)$ é a probabilidade do documento não ser relevante.

Através de regras de logaritmo e álgebra, temos que:

$$sim_p(d_j, q) = \eta \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \rho \tag{4.5}$$

onde η é a constante de normalização e ρ é baseado em estimativas de probabilidades sobre os termos (veja Seção 2.1.3 e [1] para maiores detalhes).

Modelo Funcional Probabilístico

A representação do Modelo Probabilístico na estrutura funcional é denotado por Ψ_p, onde $\Psi_p = \langle \{df_{p1}, ..., df_{pn}\}, \{qf_p\}, \Delta_p \rangle$ é um modelo funcional. O modelo probabilístico pode ser representado na estrutura funcional por:

- $df_{pj} = \{g_j\}$. Os documentos funcionais são conjuntos unários que contêm a função peso para o documento d_j. A função g_j define o peso $w_{i,j}$ no modelo probabilístico, onde $g_j(k_i) = w_{i,j}$;

- $qf_p = \{g_q\}$. As consultas funcionais são conjuntos unários que contêm a função peso da consulta q. No modelo funcional probabilístico, a função g_q define o peso de cada termo na consulta. Logo, $g_q(k_i) = w_{i,q}$;

- Função similaridade é dada por

$$\Delta_p(df_{pj}, qf_p) = \eta \sum_{i=1}^{t} g_q(k_i) \times g_j(k_i) \times \rho \tag{4.6}$$

Neste caso, considerando a constante de normalização, a função de similaridade satisfaz as propriedades de normalização, reflexividade e simetria. A constante de normalização η deve ser definida de tal forma que a função Δ retorne um valor entre 0 e 1 ($0 \leq \Delta_p(df_{pj}, qf_p) \leq 1$) e que seu maior valor (1) ocorra quando $g_q(k_i)$ for igual à $g_j(k_i)$. Dessa forma, as propriedades de normalização e de reflexividade são válidas. A propriedade de simetria também é verdadeira, pois $\Delta_p(df_{pj}, qf_p) = \Delta_p(qf_p, df_{pj})$.

4.3 Comparação e Construção de Modelos usando a Estrutura Funcional

Nesta seção, um novo modelo gerado com a utilização da estrutura funcional é discutido. Projetamos um novo modelo baseado em distância e comparamos com o modelo vetorial mostrando que eles são equivalentes.

4.3.1 Modelo Baseado em Distância

Propomos neste trabalho um Modelo Baseado em Distância para recuperação de informação e o representamos na estrutura funcional. Este modelo é baseado na distância de *Minkowski* [35, 58].

O modelo baseado em distância é um tipo de modelo espacial. Então, de modo similar ao modelo vetorial, consultas e documentos são representados como vetores em um espaço t-dimensional, onde t é o número de termos da coleção. O conjunto resposta é um *ranking* de documentos baseados na operação de distância entre o ponto que representa o documento normalizado e o ponto que representa a consulta normalizada como mostrado na Figura 4.3. Esta função de distância define o grau de similaridade entre eles.

Uma consulta q e um documento d_j são vetores compostos pelos pesos associados aos termos da

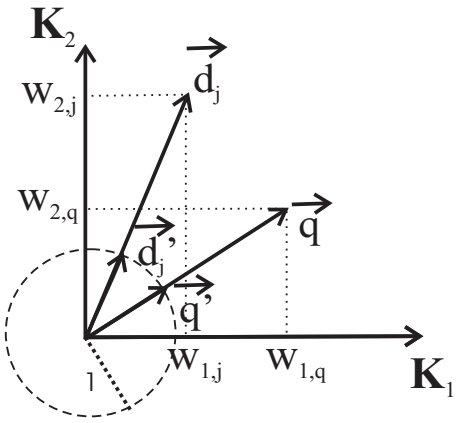

Fig. 4.3: Representação do documento $\vec{d_j}$ e da consulta $\vec{q}$ em um espaço bi-dimensional e seus vetores normalizados $\vec{d'_j}$ e $\vec{q'}$

consulta q e do documento d_j, definidos por $w_{i,j}$ e $w_{i,q}$ que são os pesos associados aos termos k_i no documento d_j e na consulta q, respectivamente.

Definimos a fórmula da distância de *Minkowski* entre o documento d_j e a consulta q como ($p \geq 1$):

$$D(\vec{d_j}, \vec{q}) = \sqrt[p]{\sum_{i=1}^{t} \left| \frac{w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,q}^2}} - \frac{w_{i,j}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2}} \right|^p} \tag{4.7}$$

Se $p = 2$, temos que a função $D(\vec{d_j}, \vec{q})$ é chamada de distância *Euclidiana* e se $p = 1$, temos que a função $D(\vec{d_j}, \vec{q})$ é chamada de distância de *Manhattan*. As funções de distância ou métrica satisfazem três propriedades: positividade ($D(\vec{q}, \vec{d_j}) \geq 0$), simetria ($D(\vec{q}, \vec{d_j}) = D(\vec{d_j}, \vec{q})$) e a desigualdade triangular mostrada abaixo. Note que a distância de *Minkowski* satisfaz a propriedade de desigualdade triangular [68]:

- $D(\vec{q}, \vec{d_j}) + D(\vec{d_j}, \vec{d_k}) \geqslant D(\vec{q}, \vec{d_k})$ (desigualdade triangular)

Mas esta função de distância não satisfaz a propriedade de reflexividade da estrutura funcional, pois a distância de um vetor com ele mesmo é igual a zero ($D(\vec{d_j}, \vec{d_j}) = 0$).

A ordenação da fórmula de distância de *Minkowski* é feita de forma crescente, pois quanto mais próximo a consulta e o documento, menor é a distância entre eles. Mas a ordenação da fórmula da similaridade é decrescente. Então, para definir o modelo baseado em distância consideramos a maior distância entre um documento e uma consulta menos a distância de *Minkowski*. Apresentamos no

próximo tópico a representação do modelo baseado em distância na estrutura funcional.

4.3.2 Modelo Funcional Baseado em Distância

Para representar o modelo baseado em distância na estrutura funcional, definimos um modelo funcional Ψ_d, onde $\Psi_d = \langle\{df_{d1}, ..., df_{dn}\}, \{qf_d\}, \Delta_d\rangle$. O modelo baseado em distância pode ser representado definindo:

- $df_{dj} = \{g_j\}$. Um documento funcional é um conjunto unário que contém a função peso para o documento d_j. A função g_j define o peso $w_{i,j}$ no modelo baseado em distância, isto é, $g_j(k_i) = w_{i,j}$;

- $qf_d = \{g_q\}$. As consultas funcionais são conjuntos unários que contêm a função peso para a consulta q. No modelo baseado em distância, a função que define o peso de cada termo na consulta é $w_{i,q}$. Logo, $g_q(k_i) = w_{i,q}$;

- A função similaridade é dada por

$$\Delta_d(df_{dj}, qf_d) = \frac{1}{\sqrt[p]{t}} \times \left(\sqrt[p]{t} - \sqrt[p]{\sum_{i=1}^{t} \left| \frac{g_q(k_i)}{\sqrt{\sum_{i=1}^{t} g_q(k_i)^2}} - \frac{g_j(k_i)}{\sqrt{\sum_{i=1}^{t} g_j(k_i)^2}} \right|^p} \right) \tag{4.8}$$

O modelo baseado em distância calcula o grau de similaridade entre um documento d_j e uma consulta q como o inverso da distância de *Minkowski* dado pela Fórmula (4.7). Consideramos a similaridade como a maior distância entre um documento e uma consulta menos a distância de *Minkowski*. Assim, quanto menor a distância maior será a similaridade. Os pesos dos termos em um documento e uma consulta são positivos e os vetores do documento e da consulta estão no primeiro quadrante ou quadrante positivo. Logo, temos que a maior distância entre um documento e uma consulta ou entre dois documentos ocorre quando eles estão ortogonais e é dada por $\sqrt[p]{t}$, onde t é o número de termos da coleção ou número de dimensões no espaço do modelo baseado em distância. O valor $\sqrt[p]{t}$ é utilizado para normalizar a distância (Fórmula 4.7) e obter a similaridade (Fórmula 4.8).

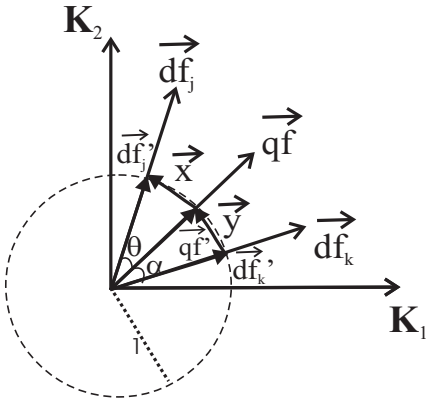

Fig. 4.4: Representação dos documentos $\vec{df}_j$ e $\vec{df}_k$, da consulta $\vec{qf}$ em um espaço bi-dimensional , seus vetores normalizados, $\vec{df}'_j$, $\vec{df}'_k$ e $\vec{qf}'$, e a distância entre eles

As representações dos documentos e consultas funcionais do modelo funcional vetorial e do modelo funcional baseado em distância são similares. Isso ocorre porque os modelos possuem o mesmo método para cálculo dos pesos dos documentos e consultas.

Através da transformação da função de distância (Eq. 4.7) para a função similaridade do modelo baseado em distância (Eq. 4.8), temos que a última satisfaz as propriedades de normalização, reflexividade e simetria. A propriedade da normalização para a função de similaridade é válida, pois com a adição da constante de normalização $\frac{1}{\sqrt[p]{t}}$ temos que $0 \leq \Delta_d(df_{dj}, qf_d) \leq 1$, visto que a maior distância entre um documento e uma consulta ou entre dois documentos é dada por $\sqrt[p]{t}$. A propriedade da reflexividade é válida, pois temos que pela distância de *Minkowski* $D(\vec{d_j}, \vec{d_j}) = 0$, então se substituirmos na Equação 4.8, podemos verificar que $\Delta_d(df_{dj}, df_{dj}) = 1$. E a propriedade de simetria também é verdadeira, pois $\Delta_d(df_{dj}, qf_d) = \Delta_d(qf_d, df_{dj})$.

4.3.3 Equivalência entre os Modelos Funcionais que representam o Modelo Vetorial e o Modelo Baseado em Distância

Comparamos os modelos vetorial e baseado em distância mostrando que eles são equivalentes.

Sejam $\Psi_v = \langle \{df_{v1}, ..., df_{vn}\}, \{qf_v\}, \Delta_v \rangle$ e $\Psi_d = \langle \{df_{d1}, ..., df_{dn}\}, \{qf_d\}, \Delta_d \rangle$ os modelos funcionais vetorial e basedo em distância, respectivamente conforme apresentamos anteriormente. A representação de um documento funcional df_j para o modelo funcional vetorial é idêntica à representação de um documento funcional para o modelo funcional baseado em distância. Então, existe uma

função bijetora $\phi_1 : \{df_{v1}, \ldots, df_{vn}\} \rightarrow \{df_{d1}, \ldots, df_{dn}\}$ que mapea um documento de Ψ_v em Ψ_d. Esta função é a função identidade. Mostramos a seguir que para toda consulta qf, se $\phi_1(df_{vj})$ = df_{dj} e $\phi_1(df_{vk})$ = df_{dk}, então as duas propriedades de equivalência (veja Definição 11) são satisfeitas.

A Figura 4.4 mostra o documento df_j que representa os documentos df_{vj} e df_{dj}, e o documento df_k que representa os documentos df_{vk} e df_{dk}. A figura também mostra os vetores normalizados, $\overrightarrow{df'_j}$ e $\overrightarrow{df'_k}$, e os vetores $\vec{x}$ e $\vec{y}$, onde $|\overrightarrow{x}|$ é a distância de *Minkowski* entre a consulta normalizada qf e o documento normalizado df_{dj} e $|\overrightarrow{y}|$ é a distância de *Minkowski* entre a consulta normalizada qf e o documento normalizado df_{dk}.

A primeira propriedade para equivalência é:

- $\Delta_v(qf, df_{vj}) = \Delta_v(qf, df_{vk}) \Leftrightarrow \Delta_d(qf, df_{dj}) = \Delta_d(qf, df_{dk})$

($\Rightarrow$) Se $\Delta_v(qf, df_{vj}) = \Delta_v(qf, df_{vk})$, então $\cos(\theta) = \cos(\alpha)$, como $0^o \leq \theta, \alpha \leq 90^o$, e $\theta = \alpha$. Isto implica que $|\overrightarrow{x}| = |\overrightarrow{y}|$, que é a distância de $Minkowski$. Logo, temos que $\Delta_d(qf, df_{dj}) = \Delta_d(qf, df_{dk})$.

($\Leftarrow$) Se $\Delta_d(qf, df_{dj}) = \Delta_d(qf, df_{dk})$, então $|\overrightarrow{x}| = |\overrightarrow{y}|$. Logo, $\theta = \alpha$. Isto implica que $\cos(\theta) = \cos(\alpha)$ e temos que $\Delta_v(qf, df_{vj}) = \Delta_v(qf, df_{vk})$.

A segunda propriedade é:

- $\Delta_v(qf, df_{vj}) > \Delta_v(qf, df_{vk}) \Leftrightarrow \Delta_d(qf, df_{dj}) > \Delta_d(qf, df_{dk})$

($\Rightarrow$) Analogamente, se $\Delta_v(qf, df_{vj}) > \Delta_v(qf, df_{vk})$, então $\cos(\theta) > \cos(\alpha)$ e $\theta < \alpha$ (consideramos $0^o \leq \theta, \alpha \leq 90^o$). Isto implica que $|\overrightarrow{x}| < |\overrightarrow{y}|$. Então, temos que $\Delta_d(qf, df_{dj}) > \Delta_d(qf, df_{dk})$.

($\Leftarrow$) Se $\Delta_d(qf, df_{dj}) > \Delta_d(qf, df_{dk})$ é verdadeiro, isto é $|\overrightarrow{x}| < |\overrightarrow{y}|$, então $\theta < \alpha$. Isto implica que $\cos(\theta) > \cos(\alpha)$. Logo, temos que $\Delta_v(qf, df_{vj}) > \Delta_v(qf, df_{vk})$.

Concluímos então que os modelos Ψ_v e Ψ_d são equivalentes. A similaridade calculada usando a medida do cosseno e a similaridade baseada na métrica de *Minkowski* para os modelos geram o mesmo *ranking*.

Neste capítulo apresentamos o meta-modelo funcional e suas definições para representação e comparação de equivalência entre modelos de RI. Também mostramos aplicações da estrutura funcional através da representação dos modelos clássicos na estrutura funcional e da construção de um modelo baseado em distância mostrando sua equivalência ao modelo clássico vetorial. No capítulo seguinte, mostramos mais aplicações da estrutura funcional em modelos de RI que combinam múltiplas evidências no contexto de redes bayesianas e vetorial.

Capítulo 5

Combinação de Múltiplas Evidências usando a Estrutura Funcional

Neste capítulo apresentamos dois estudos de caso do uso da estrutura funcional para o estudo de modelos de RI que combinam múltiplas evidências. Utilizamos a estrutura funcional para construção de novos modelos de RI que combinam múltiplas evidências no contexto de redes bayesiana e do modelo vetorial. Os modelos de redes bayesianas e vetorial para combinar múltiplas evidências são representados na estrutura funcional.

5.1 Estudo de Caso 1: Do Modelo de Redes Bayesianas para o Modelo Vetorial

No primeiro estudo de caso partimos de uma representação de redes bayesianas na estrutura funcional. Esta representação de redes bayesianas para combinar múltiplas evidências é uma extensão do modelo de redes bayesianas proposto em [33]. A partir do modelo funcional obtido, encontramos o modelo vetorial equivalente. O resultado é a representação de redes bayesianas para combinar múltiplas evidências em um modelo vetorial equivalente.

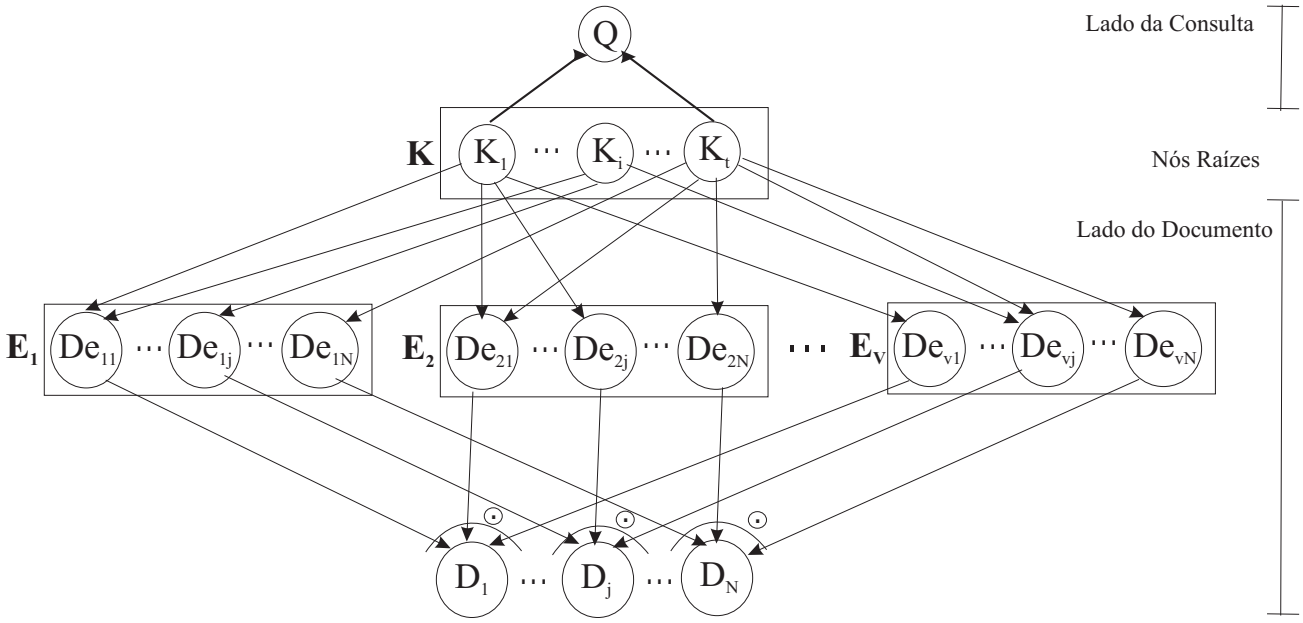

Fig. 5.1: Modelo de rede de crença para combinar múltiplas fontes de evidências

5.1.1 Modelo de Redes de Crença para Combinar Múltiplas Fontes de Evidências

O modelo de redes bayesianas discutido na Seção 2.2.2 pode ser usado para combinar múltiplas fontes de evidências, tais como evidências baseadas em palavras-chaves associadas com o conteúdo dos documentos, o texto do conteúdo de *links* e a informação da análise de *links* entre documentos da coleção. Isto pode ser obtido através da adição de novas arestas, nós e probabilidades à rede bayesiana original apresentada na Figura 2.5. O modelo extendido pode ser observado na Figura 5.1.

Propomos neste trabalho um modelo genérico de redes bayesianas para combinar múltiplas fontes de evidências. Este modelo é uma extensão do modelo de redes de crença proposto em [33]. A diferença é que generalizamos o operador usado na composição de documentos. Em [33] o operador é de disjunção, aqui ele é genérico e representado por $\odot$. A Figura 5.1 ilustra esta rede bayesiana generalizada para combinar múltiplas evidências.

Na rede bayesiana da Figure 5.1, o nó Q modela a consulta do usuário e o conjunto de nós $\mathbf{K}$ modela o conjunto de palavras-chave na coleção de documentos. Os conjuntos de nós $\mathbf{E}_1, \ldots, \mathbf{E}_v$ representam v evidências modeladas na rede. Os arcos ligando os nós de $\mathbf{K}$ aos documentos em $\mathbf{E}_1, \ldots, \mathbf{E}_v$ indicam que os termos da consulta induzem crença aos nós de documentos de acordo com

a evidência representada, por exemplo, *links*. Para representar uma nova fonte de evidência e_i nesta rede, novos nós $De_{i,j}$ são associados com cada documento D_j no conjunto resposta para a consulta Q. O conjunto de nós **K** é usado para modelar a ocorrência de termos na consulta Q que induz valores de crença em cada um dos nós dos conjuntos $\mathbf{E}_1, \ldots, \mathbf{E}_v$. O nó D_j representa a combinação de todas as evidências modeladas.

O *ranking* de um documento é calculado como a probabilidade $P(dj|q)$, como a seguir:

$$P(d_j|q) = \eta \sum_{\forall \mathbf{k}} P(d_j|\mathbf{k}) \times P(q|\mathbf{k}) \times P(\mathbf{k}) \tag{5.1}$$

onde η é uma constante de normalização. Detalhes da derivação dessa expressão podem ser encontrados na Seção 2.2 e em [54]. Entretanto, a probabilidade condicional $P(dj|\mathbf{k})$ agora depende de múltiplas evidências combinadas através do operador $\odot$, que pode ser os operadores disjuntivo, conjuntivo e noisy-OR.

Para o operador disjuntivo, isto é realizado através da equação:

$$P(d_j|\mathbf{k}) = 1 - (1 - P(de_{1j}|\mathbf{k})) \times (1 - P(de_{2j}|\mathbf{k})) \times \cdots \times (1 - P(de_{vj}|\mathbf{k})) \tag{5.2}$$

onde $P(de_{ij}|\mathbf{k})$ é o valor calculado para cada evidência E_i em relação ao documento d_j que denotamos aqui como E_{ij}. E_{ij} pode ser, por exemplo, o peso da parte de conteúdo do documento d_j, calculado pelo modelo clássico vetorial, ou o grau de *hub* e *autoridade* do documento d_j. E $P(q|\mathbf{k})$ é definido por:

$$P(q|\mathbf{k}) = \begin{cases} 1, & \text{se } \mathbf{q} = \mathbf{k} \\ 0, & \text{caso contrário} \end{cases} \tag{5.3}$$

Substituindo cada $P(de_{ij}|\mathbf{k})$ por E_{ij} em Eq.(5.2), e substituindo as Eq.(5.2) e (5.3) em Eq.(5.1), definindo a probabilidade *a priori* $P(\mathbf{k})$ como constante e considerando que a constante η não influencia no resultado final do *ranking*, podemos definir a *função similaridade* como:

$$sim(d_j, q) = 1 - (1 - E_{1j})(1 - E_{2j}) \ldots (1 - E_{vj}) \tag{5.4}$$

Observe que qualquer evidência e_i pode ser ignorada, atribuindo E_{ij} = 0. Note que esta função de similaridade não satisfaz a propriedade de simetria, pois $sim(d_j, q) \neq sim(q, d_j)$.

Analogamente, para o operador conjuntivo, temos a multiplicação dos valores de cada evidência como mostrado na seguinte função:

$$sim(d_j, q) = E_{1j} \times E_{2j} \cdots \times E_{vj} \tag{5.5}$$

Note que se para qualquer evidência e_i, E_{ij} = 0, então $sim(d_j, q)$ = 0, ignorando todos as outras evidências. Por isto o operador conjuntivo não é muito utilizado na prática.

A combinação no modelo usando os operadores disjuntivo e conjuntivo não faz a hipótese *a priori* sobre a importância de cada fonte de evidência. As probabilidades a serem combinadas dependem somente das características dos algoritmos e dos parâmetros usados. Entretanto, o modelo pode ser modificado para permitir a inserção de pesos. Isto pode ser realizado utilizando o operador noisy-OR (maiores detalhes sobre este operador podem ser encontrados em [45]). Então, temos a seguinte equação para a função de similaridade:

$$sim(d_j, q) = 1 - (1 - W_1 \times E_{1j})(1 - W_2 \times E_{2j}) \ldots (1 - W_v \times E_{vj}) \tag{5.6}$$

onde $W_1 \ldots W_v$ são os pesos atribuidos para cada evidência $e_1, \ldots, e_v$, respectivamente. Estes pesos podem sem definidos pelo usuário, podem depender ou não da consulta ou podem ser automaticamente calculados.

Para simplificar nossa notação, seja R_{jq} a função de *ranking* do modelo vetorial de D_j com relação à consulta Q. A informação fornecida pelo modelo vetorial pode ser incluida como uma evidência fazendo E_{1j} = R_{jq}.

5.1.2 Modelo Funcional de Redes de Crença para Combinar Múltiplas Fontes de Evidências

A representação do modelo de redes de crença apresentado anteriormente na estrutura funcional é mostrado aqui. Para representar o modelo de redes de crença genérico que combina múltiplas fontes

de evidências usando o operador disjuntivo na estrutura funcional, definimos o modelo funcional $\Psi_{ng} = \langle\{df_{ng_1}, ..., df_{ng_n}\}, \{qf_{ng}\}, \Delta_{ng}\rangle$. O modelo bayesiano com múltiplas evidências pode ser representado na estrutura funcional por:

- $df_{ng_j} = \{g_{e1_j}, g_{e2_j}, \ldots, g_{ev_j}\}$, onde $g_{e1_j}(k_i) = w_{i,j}$ é a função peso calculada pelo modelo vetorial de cada termo do documento d_j e $g_{e2_j} \ldots g_{ev_j}$ são funções que definem valores para as evidências $e_2, \ldots, e_v$ associadas com o documento d_j,respectivamente.

- $qf_{ng} = \{g_{e1_q}, g_{e2_q}, \ldots, g_{ev_q}\}$, onde $g_{e1_q}(k_i) = w_{i,q}$ e os outros termos funcionais são definidos de forma análoga aos documentos funcionais.

- A função similaridade é dada por

$$\Delta_{ng}(df_{ng_j}, qf_{ng}) = 1 - (1 - R_{j,q})(1 - g_{e2_j} g_{e2_q}) \ldots (1 - g_{ev_j} g_{ev_q}) \tag{5.7}$$

onde $R_{j,q}$ é dado pela Equação 4.3.

Considerando o operador disjuntivo, representado pela Equação 5.4 satisfaz a a propriedade de normalização, considerando que os valores das evidências estejam normalizados ($0 \leq E_{1j}, \ldots, E_{vj} \leq 1$). Mas esta equação (Eq.5.4) não satisfaz as propriedades de simetria, pois $sim(d_j, q) \neq sim(q, d_j)$ e da reflexividade, pois $sim(d_j, d_j)$ pode ser diferente de 1.

Analogamente, para o operador conjuntivo, a Equação 5.5 satisfaz a propriedade de normalização, mas as propriedade de simetria e reflexividade não são satisfeitas.

Logo, como a Eq.(5.4) não satisfaz as propriedades de simetria e reflexividade, utilizamos a Eq.(5.7) que sobrepõe a Eq.(5.4) atribuindo $g_{e2_q} = \ldots = g_{ev_q} = 1$. Note que modificamos a função original de similaridade Eq.(5.4) para Eq.(5.7) visando satisfazer as propriedades necessárias para uma função de similaridade na estrutura funcional. Observe que neste caso, temos que o operador noisy-OR indroduzido anteriormente é dado pela Eq.(5.6). A função R_{jq} satisfaz as três propriedades como foi mostrado na Seção 4.2.1. Neste caso, a propriedade de normalização é satisfeita, pois os valores das funções que definem os valores das evidências são valores normalizados ($0 \leq g_{e2_j}, \ldots, g_{ev_j} \leq 1$) e a função R_{jq} também. A propriedade de reflexividade também é válida, pois se $R_{jj} = 1$, então $\Delta_{ng}(df_{ng_j}, df_{ng_j}) = 1$ e a propriedade de simetria também é válida, pois

$\Delta_{ng}(df_{ngj}, qf_{ng}) = \Delta_{ng}(qf_{ng}, df_{ngj})$. Então, para representar a rede de crença usando o operador disjuntivo precisamos modificar a função similaridade para o operador noisy-OR para satisfazer as propriedades da estrutura funcional.

A alteração realizada na função de similaridade pode ser realizada devido ao poder de abstração das funções. Isto é importante, por exemplo, para trabalhar com a clusterização de documentos e neste caso temos que calcular a similaridade entre dois documentos.

5.1.3 Modelo Vetorial para Combinar Múltiplas Fontes de Evidências

Definimos um modelo vetorial extendido para combinar múltiplas evidências equivalente à rede bayesiana anterior. No modelo clássico vetorial, o conjunto de termos $\{k_i | 1 \leq i \leq t\}$ formam os eixos do modelo vetorial. Os documentos e consultas são representados como vetores no espaço: $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$ e $q = (w_{1q}, w_{2q}, \ldots, w_{tq})$, respectivamente.

Propomos neste trabalho um modelo que combina informações de múltiplas fontes de evidências através de uma extensão do modelo de espaço vetorial. Para isso, extendemos o espaço vetorial adicionando $v - 1$ novos eixos, onde $v - 1$ é o número de novas evidências. Figure 5.2 mostra este modelo vetorial para combinação de múltiplas fontes de evidências.

Neste caso, a equação da função de similaridade é:

$$sim(d_j, q) = 1 - (1 - R_{jq})(1 - E_{2q} \times E_{2j}) \ldots (1 - E_{2v} \times E_{vj}) \tag{5.8}$$

onde R_{jq} é calculado pelo cosseno do modelo vetorial, $E_{2q}, \ldots, E_{vq}$ são os valores de cada evidência $e_2, \ldots, e_v$ associado à consulta q e $E_{2j}, \ldots, E_{vj}$ são os valores de cada evidência $e_2, \ldots, e_v$ associados ao documento d_j, respectivamente.

A seguir representamos este modelo na estrutura funcional e discutimos suas propriedades.

5.1.4 Modelo Funcional Vetorial para Combinar Múltiplas Fontes de Evidências

Representamos o modelo vetorial para combinar múltiplas fontes de evidências apresentado anteriormente na estrutura funcional. Para representar o modelo genérico vetorial para combinação de

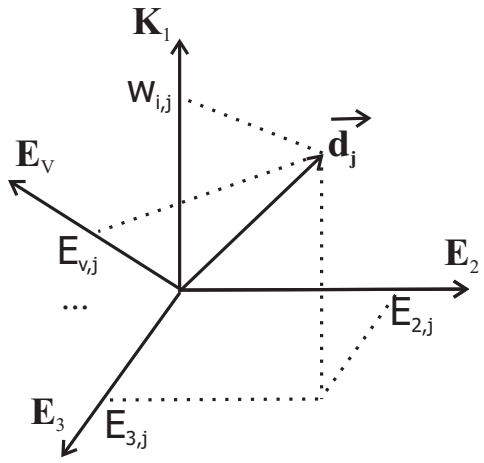

Fig. 5.2: Modelo genérico vetorial para combinação de múltiplas fontes de evidências

múltiplas evidências na estrutura funcional, definimos o modelo funcional $\Psi_{vg} = \langle\{df_{vg_1}, ..., df_{vg_n}\}, \{qf_{vg}\}, \Delta_{vg}\rangle$, onde:

- $df_{vg_j} = \{g_{e1_j}, g_{e2_j}, \ldots, g_{ev_j}\}$, onde g_{e1_j} = $w_{i,j}$ é a função peso calculada pelo modelo vetorial de cada termo do documento d_j e $g_{e2_j} \ldots g_{ev_j}$ são funções que definem valores para as evidências $e_2, \ldots, e_v$ associadas com o documento d_j,respectivamente.

- $qf_{vg} = \{g_{e1_q}, g_{e2_q}, \ldots, g_{ev_q}\}$, onde $g_{e1_q} = w_{i,j}$ e os outros termos são definidos de forma análoga aos documentos funcionais.

- Função similaridade é dada por

$$\Delta_{vg}(df_{vg_j}, qf_{vg}) = 1 - (1 - R_{jq})(1 - g_{e2_j} g_{e2_q}) \ldots (1 - g_{ev_j} g_{ev_q}) \tag{5.9}$$

Projetamos um modelo vetorial para combinar múltiplas fontes de evidências equivalente ao modelo de redes de crença para combinar múltiplas evidências. Verificamos que os modelos funcionais correspondentes são equivalentes por construção. Logo, existe a função bijetora identidade ϕ e as duas propriedades de equivalência são satisfeitas, pois possuem a mesma função de similaridade. Os modelos Ψ_{vg} e Ψ_{ng} são equivalentes, e geram o mesmo *ranking*.

5.2 Estudo de Caso 2: Do Modelo Vetorial para o Modelo de Redes Bayesianas

O meta-modelo funcional representa modelos de RI em uma linguagem funcional. Através dessa representação podemos pensar em outras alternativas de combinação de múltiplas evidências como a apresentada nesta seção. A correspondência entre a combinação de múltiplas evidências nas redes bayesianas e no modelo vetorial utiliza a estrutura funcional como um unificador de modelos.

Esta seção apresenta um segundo estudo de caso para combinar múltiplas fontes de evidências em um modelo vetorial e determina o modelo de redes de crença equivalente com a utilização da estrutura funcional. Propomos uma outra forma de combinar fontes de evidências no modelo vetorial. Representamos este modelo vetorial extendido na estrutura funcional e encontramos o modelo de redes de crença equivalente para combinar múltiplas evidências.

5.2.1 Modelo Vetorial para Combinar Múltiplas Fontes de Evidências

Podemos combinar múltiplas evidências no modelo vetorial modificando a função de similaridade. Outra forma de combinar fontes de evidências no modelo vetorial é usar a função de similaridade do cosseno.

A modelagem vetorial é a mesma apresentada na Seção 5.1.3. Este modelo também é uma extensão do modelo vetorial através da adição de $v-1$ novos eixos, onde $v-1$ é o número de novas evidências. Este modelo vetorial extendido para combinar múltiplas evidências é apresentado na Figura 5.2. Além dos eixos $\mathbf{K}_1, \ldots, \mathbf{K}_t$ que representam os termos do vocabulário ou a evidência conteúdo (evidência $\mathbf{E}_1$), foram inseridos os eixos $\mathbf{E}_2, \ldots, \mathbf{E}_v$ representando as $v-1$ novas evidências.

A *função de similaridade* é definida por:

$$sim(d_j, q) = \frac{(\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}) + E_{2j}E_{2q} + \cdots + E_{vj}E_{vq}}{\sqrt{(\sum_{i=1}^{t} w_{i,j}^2) + E_{2j}^2 + \cdots + E_{vj}^2} \times \sqrt{(\sum_{i=1}^{t} w_{i,q}^2) + E_{2q}^2 + \cdots + E_{vq}^2}} \tag{5.10}$$

onde $w_{i,j}$ é o peso do termo k_i no documento d_j, $w_{i,q}$ é o peso do termo k_i na consulta q, $E_{2j}, \ldots, E_{vj}$ são os valores de cada evidência $e_2, \ldots, e_v$ associados ao documento d_j e $E_{2q}, \ldots, E_{vq}$ são os valores de cada evidência $e_2, \ldots, e_v$ associada à consulta q, respectivamente.

5.2.2 Modelo Vetorial Funcional para Combinar Múltiplas Fontes de Evidências

Representamos o modelo vetorial que combina múltiplas evidências apresentado anteriormente na estrutura funcional. Para representar o modelo genérico vetorial que combina múltiplas fontes de evidências na estrutura funcional definimos o modelo funcional $\Psi_{vc} = \langle\{df_{vc_1}, ..., df_{vc_n}\}, \{qf_{vc}\}, \Delta_{vc}\rangle$. O modelo vetorial com múltiplas fontes de evidências pode ser representado na estrutura funcional por:

- $df_{vc_j} = \{g_j, g_{e2_j}, \ldots, g_{ev_j}\}$, onde $g_j(k_i) = w_{i,j}$ é a função que define o peso dos termos no documento e $g_{e2_j} \ldots g_{ev_j}$ são funções que definem valores associados às evidências $e_1, \ldots, e_v$, respectivamente.

- $qf_{vc} = \{g_q, g_{e2_q}, \ldots, g_{ev_q}\}$, onde os termos funcionais são definidos de forma análoga aos documentos funcionais.

- Função de similaridade é dada por

$$\Delta_{vc}(df_{vc_j}, qf_{vc}) = \frac{(\sum_{i=1}^{t} g_j(k_i) \cdot g_q(k_i)) + g_{e2_j} g_{e2_q} + \cdots + g_{ev_j} g_{ev_q}}{\sqrt{(\sum_{i=1}^{t} g_j(k_i)^2) + g_{e2_j}^2 + \cdots + g_{ev_j}^2} \times \sqrt{(\sum_{i=1}^{t} g_q(k_i)^2) + g_{e2_q}^2 + \cdots + g_{ev_q}^2}} \tag{5.11}$$

Note que esta função similaridade satisfaz as propriedades de normalização, reflexividade e simetria. A propriedade de normalização é válida, pois como a função de similaridade é a função cosseno entre dois vetores que contêm coordenadas positivas (se encontram no primeiro quadrante), então retorna um valor entre 0 e 1 ($0 \leq cos(\vec{a}, \vec{b}) \leq 1$). A propriedade reflexividade é válida, pois se $cos(\vec{a}, \vec{a}) = 1$, então $\Delta_{vc}(df_{vc_j}, df_{vc_j}) = 1$ e a propriedade da simetria também é válida, pois $cos(\vec{a}, \vec{b}) = cos(\vec{b}, \vec{a})$, logo $\Delta_{vc}(df_{vc_j}, qf_{vc}) = \Delta_{vc}(qf_{vc}, df_{vc_j})$.

5.2.3 Modelo de Redes de Crença para Combinar Múltiplas Fontes de Evidências

Desenvolvemos um modelo de redes de crença para combinar múltiplas evidências equivalente ao modelo vetorial para combinar múltiplas fontes de evidências usando a função de similaridade do cosseno. A rede resultante é mostrada na Figura 5.3.

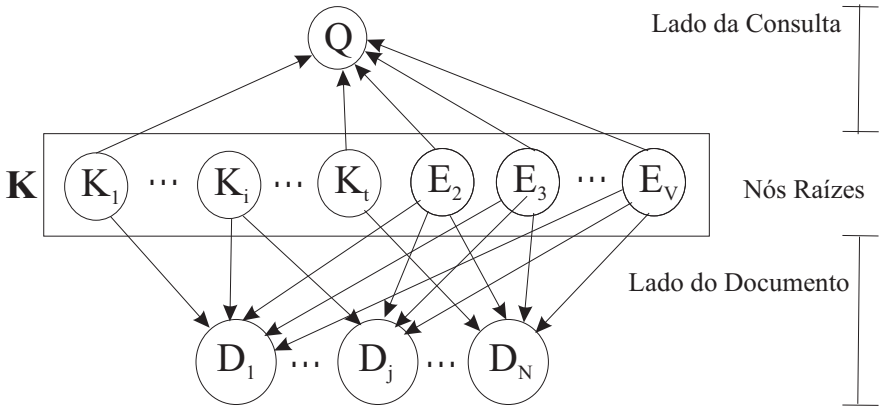

Fig. 5.3: Modelo genérico de redes de crença para combinar múltiplas fontes de evidências

Neste modelo de redes bayesianas, o *ranking* é calculado como:

$$sim(d_j, q) = \frac{(\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}) + E_{2j}E_{2q} + \cdots + E_{vj}E_{vq}}{\sqrt{(\sum_{i=1}^{t} w_{i,j}^2) + E_{2j}^2 + \cdots + E_{vj}^2} \times \sqrt{(\sum_{i=1}^{t} w_{i,q}^2) + E_{2q}^2 + \cdots + E_{vq}^2}} \tag{5.12}$$

A derivação desta equação é realizada de forma similar às apresentadas nas Seções 2.2.2 e 2.2.3 e em [54] para Eq.(5.1) fazendo as substituições necessárias para modelagem da função cosseno.

5.2.4 Modelo Funcional de Redes de Crença que Combina Múltiplas Fontes de Evidências

A representação do modelo de redes de crença apresentado anteriormente na estrutura funcional é mostrado aqui. Para representar o segundo modelo de redes bayesianas com múltiplas fontes de evidências na estrutura funcional definimos o modelo funcional $\Psi_{nc} = \langle \{df_{nc_1}, ..., df_{nc_n}\}, \{qf_{nc}\}, \Delta_{nc} \rangle$, onde:

- $df_{nc_j} = \{g_j, g_{e2_j}, \ldots, g_{ev_j}\}$, onde $g_j(k_i) = w_{i,j}$, é função que define o peso dos termos no documento e $g_{e2_j} \ldots g_{ev_j}$ são funções que definem valores associados às evidências $e_1, \ldots, e_v$, respectivamente.

- $qf_{nc} = \{g_q, g_{e2_q}, \ldots, g_{ev_q}\}$, onde os termos funcionais são definidos de forma análoga aos documentos funcionais.

- A função de similaridade é dada por

$$\Delta_{nc}(df_{nc_j}, qf_{nc}) = \frac{(\sum_{i=1}^{t} g_j(k_i) \cdot g_q(k_i)) + g_{e2_j} g_{e2_q} + \cdots + g_{ev_j} g_{ev_q}}{\sqrt{(\sum_{i=1}^{t} g_j(k_i)^2) + g_{e2_j}^2 + \cdots + g_{ev_j}^2} \times \sqrt{(\sum_{i=1}^{t} g_q(k_i)^2) + g_{e2_q}^2 + \cdots + g_{ev_q}^2}} \tag{5.13}$$

Propomos um modelo de redes de crença para combinar múltiplas evidências equivalente ao modelo vetorial para combinar múltiplas evidências com a função de similaridade do cosseno. Podemos verificar que os modelos funcionais Ψ_{vc} e Ψ_{nc} são equivalentes pela sua própria construção.

Neste capítulo analisamos a combinação de múltiplas evidências apresentando aplicações do uso da estrutura funcional para combinar múltiplas evidências nos contextos de redes bayesianas e modelo de espaço vetorial. No primeiro estudo de caso encontramos um modelo vetorial correspondente à um modelo de redes bayesianas extendido e no segundo estudo de caso construímos um modelo de redes bayesianas equivalente ao mesmo modelo vetorial do primeiro estudo de caso modificando apenas a função de similaridade para a função do cosseno. Através desses estudos podemos verificar que a combinação de múltiplas evidências na rede bayesiana de crença pode ser realizada de várias formas, sendo que cada uma corresponde à uma função de similaridade no modelo vetorial. A análise dessa correspondência é feita através da estrutura funcional. Com isso, mostramos que o meta-modelo funcional permite-nos desenvolver novos modelos e ajuda desenvolvedores a modificar esses modelos para extendê-los com novas fontes de evidências.

Capítulo 6

Proposta de uma Ferramenta para Avaliação de Desempenho de SRI

Neste capítulo apresentamos as idéias iniciais de uma aplicação da estrutura funcional para comparação experimental entre modelos de RI. Trata-se da utilização do meta-modelo funcional para construção de uma ferramenta para avaliação experimental de sistemas de recuperação de informação.

Sistemas de recuperação de informação têm sido avaliados e comparados experimentalmente há vários anos. Conhecer a eficiência de sistemas de recuperação de informação é importante não só para os pesquisadores de RI, mas também para quem usa estes sistemas. Pesquisadores e usuários precisam ter maneiras efetivas para saber quão bons são os sistemas para uma dada tarefa. Na Seção 6.1 apresentamos uma introdução sobre a proposta para construção de uma ferramenta para avaliação de desempenho entre modelos de RI utilizando os conceitos da estrutura funcional e nas Seções 6.2, 6.3, 6.4, 6.5 e 6.6 descrevemos os componentes da ferramenta.

6.1 Introdução

A estrutura funcional representa modelos de RI em uma mesma linguagem: a linguagem funcional. Isto torna prático a implementação de modelos de RI para comparação experimental. A idéia é propor uma ferramenta, baseada nos conceitos da estrutura funcional, como um sistema para avaliação de

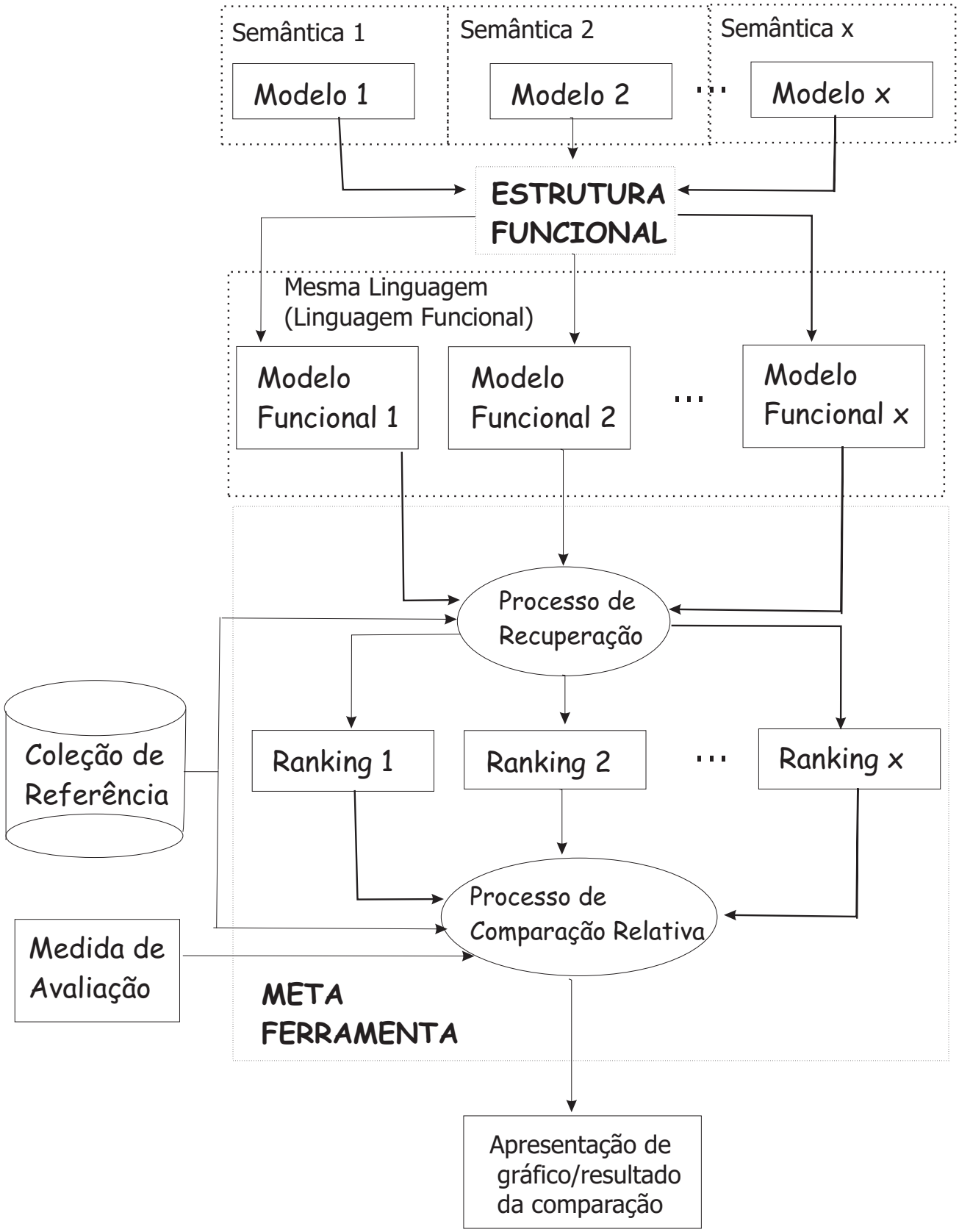

Fig. 6.1: Representação gráfica da proposta da ferramenta para avaliação de desempenho entre modelos de RI

desempenho entre modelos de RI que permita a comparação relativa entre modelos de RI através de experimentos.

A Figura 6.1 mostra uma representação gráfica do funcionamento da ferramenta para avaliação de resultados entre modelos de RI. A ferramenta deve possuir uma interface de interação com o usuário, permitindo a definição dos modelos a serem comparados, a escolha da medida de avaliação de desempenho e da coleção de referência a ser utilizada. Os componentes da ferramenta incluem a especificação dos modelos funcionais a serem comparados , a especificação da coleção de referência, da medida de avalição, o processo de recuperação e o processo de comparação relativa que apresenta o resultado da comparação. Descrevemos a seguir os componentes da ferramenta.

6.2 Especificação dos Modelos Funcionais

Para calcular o *ranking*, um SRI adota um modelo de RI para representar os documentos e as consultas. Na ferramenta, os modelos de RI devem ser passados para uma linguagem funcional através da estrutura funcional. Então, o primeiro passo para a avaliação experimental de resultados usando a ferramenta é a especificação dos modelos funcionais a serem comparados.

Um modelo funcional é definido pela tupla

$$\Psi = \langle \mathbf{D}_f, \mathbf{Q}_f, \mathbf{T}_f, \Delta \rangle$$

- $\mathbf{D}_f$ é o conjunto de documentos funcionais $\{df_1, \ldots, df_n\}$
- $\mathbf{Q}_f$ é o conjunto finito de consultas funcionais $\{qf_1, \ldots, qf_m\}$
- $\mathbf{T}_f$ é o conjunto de termos funcionais dos documentos e das consultas funcionais $\{g_1, \ldots, g_v\}$
- Δ é a uma função similaridade, que satifaz três propriedades, a saber: normalização, reflexividade e simetria.

O processo de especificação dos modelos funcionais envolve a especificação dos documentos funcionais $\mathbf{D}_f$, das consultas funcionais $\mathbf{Q}_f$, dos termos funcionais para os documentos e consultas funcionais $\mathbf{T}_f$ e da função de similaridade Δ que satisfaz as propriedades de normalização, reflexividade e simetria.

Então, um usuário da ferramenta deve definir para cada modelo funcional $\Psi_1, \ldots, \Psi_x$:

- Nome do modelo
- Definição das Consultas Funcionais ($\mathbf{Q}_f$):
  - N_{qf}: número de termos funcionais para uma consulta
  - Termos funcionais para as consultas funcionais: conjunto de funções cuja semântica relaciona os termos de uma consulta
- Definição dos Documentos Funcionais ($\mathbf{D}_f$):
  - N_{df}: número de termos funcionais para um documento
  - Termo funcional para os documentos funcionais: conjunto de funções cuja semântica relaciona os termos de um documento
- A função similaridade Δ

Os termos funcionais das consultas e documentos funcionais e a função similaridade são especificadas através de uma linguagem funcional utilizada pela ferramenta.

A ferramenta deve ser capaz de processar o modelo funcional a partir desta especificação. Isto será descrito no processo de recuperação na Seção 6.5.

6.3 Especificação da Coleção de Referência

Um sistema de recuperação de informação pode ser avaliado através de consultas que fazem parte de uma coleção de referência. Uma coleção de referência é formada por um conjunto de documentos, um conjunto de consultas e um conjunto de documentos relevantes para cada consulta geralmente classificados por usuários especialistas nos temas envolvidos.

Um exemplo de coleção de referência é a conhecida coleção TIPSTER, usada na Text REtrieval Conference (TREC), descrita em [28]. A TIPSTER é uma coleção de cerca de um milhão de documentos, obtidos de várias fontes, tais como o *Wall Street Journal*. Nesta coleção há um conjunto de consultas e para cada consulta é fornecido um conjunto ideal de documentos resposta.

O processo de avaliação de desempenho ou comparação relativa e o processo de recuperação são baseados em uma coleção de referência. A especificação da coleção de referência envolve a seleção

de uma coleção pelo usuário da ferramenta, por exemplo: WBR-99 *Collection*, TREC (*Text Retrieval Conferences*), CACM (*Communications of the ACM*), CISI (*Collection of Institute of Scientific Information*), CFC (*Cystic Fibrosis Collection*), entre outras coleções que podem ser mapeadas no sistema e a especificação do intervalo das consultas a serem testadas (N_q), por exemplo, $2..20$ significa que devem ser avaliadas as consultas de números 2 à 20.

Uma coleção de referência possui um conjunto de arquivos contendo informações que especificam partes dos documentos coletados. A ferramenta deve possuir o caminho de todos os arquivos utilizados na realização dos testes e seus arquivos indexados parametrizado para cada coleção de referência. Cada coleção mapeada na ferramenta deverá possuir estruturas de dados que foram previamente indexadas, tais como o vocabulário indexado, o vetor de termos, a lista invertida ou arquivo invertido indexado, o arquivo de normas dos documentos, o arquivo de IDF dos termos, o arquivo de consultas da coleção, o arquivo de documentos relevantes para cada consulta [23] e o arquivo de *links* para as coleções que permitem o cálculo de *links* (hub e autoridade). Descrevemos a seguir algumas dessas estruturas.

O vocabulário é o conjunto de palavras ou termos distintos que ocorrem nos documentos da coleção. As palavras do vocabulário, geralmente, estão armazenadas em ordem lexicográfica. O arquivo de vocabulário pode ser dividido em dois: o arquivo de índice do vocabulário que contém um identificador para o termo e o arquivo de dados do vocabulário contendo os termos. O arquivo de vetor de termos armazena um vetor de termos para cada documento da coleção e o arquivo invertido armazena o inverso do vetor de termos, isto é, armazena uma lista de documentos para cada termo do vocabulário. O arquivo de IDF dos termos contem a freqüência inversa de cada termo da coleção e o arquivo de normas contem as normas dos documentos. O cálculo da norma e do IDF segue o padrão descrito na Seção 2.1.2. O vetor de consultas armazena o vetor de termos para algumas consultas da coleção e o vetor de respostas armazena a lista de documentos relevantes para cada consulta.

6.4 Especificação da Medida de Avaliação

A especificação da medida de avaliação é utilizada no processo de comparação relativa e envolve a seleção da medida de avaliação utilizada na comparação por parte do usuário da ferramenta. A seguir,

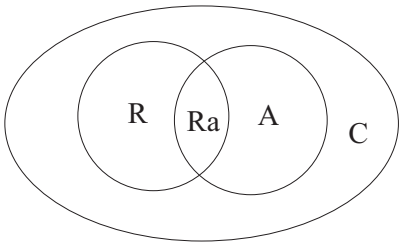

Fig. 6.2: Conjuntos para definição de precisão e revocação

descrevemos algumas medidas de avaliação utilizadas em recuperação de informação.

As medidas mais utilizadas para avaliar o desempenho de sistemas de RI são precisão e revocação e foram originalmente propostos por [36]. São medidas utilizadas para avaliar a eficácia de um sistema de RI, ou seja, elas medem a habilidade do sistema de recuperar os documentos relevantes e, ao mesmo tempo, de evitar os não relevantes [65].

As seguintes definições são necessárias para se entender esses dois conceitos. A avaliação de desempenho de SRI é realizada através da comparação entre conjunto resposta e conjunto ideal. Seja C o conjunto de documentos da coleção. Seja R o conjunto de documentos relevantes para uma dada consulta, identificado por um grupo de especialistas, e $|R|$ o número de documentos em R. Seja A o conjunto de documentos da resposta retornado pelo sistema de RI e $|A|$ o número de documentos em A. Seja Ra o conjunto de documentos relevantes do conjunto resposta A, resultado da interseção entre os conjuntos R e A, e seja $|Ra|$ o número de documentos no conjunto Ra. A Figura 6.2 ilustra esses conceitos e seus relacionamentos.

6.4.1 Precisão

Precisão é a fração de documentos recuperados que são relevantes [1], ou seja, é uma medida da capacidade do sistema de recuperar somente documentos relevantes. É calculada pela fórmula:

$$Precisao = \frac{Ra}{|R|} \tag{6.1}$$

6.4.2 Revocação

Revocação é a fração de documentos relevantes que foram recuperados [1], ou seja, é uma medida da capacidade do sistema de recuperar todos os documentos relevantes. É calculada pela fórmula:

$$Revocacao = \frac{Ra}{|A|} \tag{6.2}$$

6.4.3 Precisão nos X primeiros

A medida de precisão nos X primeiros é a precisão na posição X do *ranking*. A idéia de realizar a avaliação com base nesta medida visa focar a avaliação nos documentos efetivamente observados pelo usuário [57].

6.4.4 Precisão-R

Precisão-R é a medida que calcula a precisão na posição R do *ranking*, onde R é o tamanho do conjunto ideal.

6.4.5 Medida-E

A medida-E utiliza uma medida que combina as medidas de precisão e revocação [65], dada pela equação a seguir. A vantagem do uso da Medida-E é que ela permite dar ênfase na precisão ou na revocação.

$$E_j = 1 - \frac{1 + \beta^2}{\beta^2/Revocacao_j + 1/Precisao_j} \tag{6.3}$$

E_j é a medida-E na posição j do *ranking*. Quanto maior o valor de β maior o interesse em precisão.

As opções de medidas de avaliação para seleção na ferramenta seriam, por exemplo, Precisão e Revocação, Precisão nos X primeiros, Precisão-R e Medida-E. Se a opção selecionada fosse Precisão nos X primeiros, então o usuário deverá especificar o número X que representa o número dos

primeiros documentos retornados. Se a opção selecionada fosse Medida-E, então o usuário deverá especificar o valor da constante β que define a relativa importância de precisão e revocação.

A decisão de quais medidas utilizar em uma avaliação depende da aplicação e há discussões sobre a confiabilidade de tais medidas [62].

6.5 Processo de Recuperação

O processo de recuperação da ferramenta pode ser dividido em 3 etapas: o processo de indexação, a interpretação dos modelos funcionais e a geração dos $rankings$.

O processo de indexação envolve a criação de estruturas de dados associados à parte textual dos documentos através da análise do conteúdo dos documentos e traduzidos em termos de uma linguagem de indexação, tais como, as estruturas de arranjos de sufixos e arquivos invertidos [23]. Estas estruturas podem conter dados sobre características dos termos na coleção de documentos, tais como a freqüência de cada termo em um documento. Os índices são construídos para cada coleção de documentos e são usados para acelerar a tarefa de recuperação. A representação gerada no processo de indexação identifica o documento e define seus pontos de acesso para a recuperação dos documentos. A geração dos $rankings$ utiliza essas estruturas de dados geradas no processo de indexação.

A ferramenta deve ser capaz de interpretar os modelos funcionais, por exemplo, através de uma análise sintática e léxica da especificação dos modelos funcionais na linguagem funcional.

A terceira etapa do processo de recuperação é a geração das listas de documentos recuperados ou $rankings$ $1, \ldots, x$ para cada modelo funcional especificado $\Psi_1, \ldots, \Psi_x$, respectivamente, e para as consultas mapeadas na coleção de referência. O processo de recuperação utiliza a função de similaridade (Δ) especificada para cada modelo funcional para realizar a geração dos $rankings$. Então, os $rankings$ são classificados em ordem decrescente de acordo com o grau de similaridade entre os documentos funcionais e as consulta funcionais.

Assim, a ferramenta processa os modelos que estão especificados através da linguagem funcional da estrutura funcional e realiza o processo de busca ou geração dos $rankings$ para os modelos funcionais utilizando as estrutura de dados geradas no processo de indexação.

6.6 Processo de Comparação Relativa ou Avaliação de Resultados

A avaliação de desempenho de SRI é realizada através da comparação entre conjunto resposta ($ranking$) e conjunto ideal. Um SRI classifica os documentos recuperados para cada consulta, de acordo com uma ordem de relevância. Avalia-se o SRI através da comparação dos $rankings$ gerados por este sistema e o conjunto ideal de respostas. Para isso, o conjunto de respostas retornado pelo SRI é examinado e comparado com o conjunto ideal através das medidas de avaliação.

A comparação entre os modelos $\Psi_1, \ldots, \Psi_x$ com relação a qualidade é descrita a seguir. Dados os parâmetros informados pelo usuário, a ferramenta gerará os *rankings* para cada modelo $\Psi_1, \ldots, \Psi_x$ no processo de recuperação. Após a geração dos *rankings*, o processo de comparação deve realizar o cálculo da medida de avaliação selecionada pelo usuário baseado nos $rankings$ e no conjunto ideal das respostas que estão armazenados na coleção de referência. Finalmente, o resultado da comparação deve ser apresentado para o usuário para análise dos resultados obtidos.

Neste capítulo apresentamos as idéias de desenvolvimento de uma ferramenta genérica, para avaliação de resultados de sistemas de recuperação de informação utilizando alguns conceitos do metamodelo funcional.

Capítulo 7

Conclusões e Trabalhos Futuros

7.1 Conclusões

A estrutura funcional define um nível de abstração e fornece uma poderosa ferramenta para representar, comparar, combinar e construir modelos de RI. O nível de abstração é maior que os modelos de RI tradicionais e menor que outros meta-modelos genéricos como Caracterização BR-Formal. Isso permite trabalhar com aplicações teóricas e práticas, tornando-o prático no sentido de implementação e não tão genérico. A estrutura funcional permite a análise de diferentes modelos usando diferentes semânticas de modelagem. Este *framework* é uma simples, poderosa e flexível ferramenta de modelagem para RI.

Além disso, este *framework* é um meta-modelo para modelos de RI e oferece um poder de expressividade para representação de modelos. A representação funcional é importante para estudar características e propriedades dos modelos de RI. O *framework* pode ser usado para generalizar todos os modelos de RI que podem ser expressos por um algoritmo, pois ele é baseado em funções. Uma de nossas contribuições é a proposta de um meta-modelo capaz de formular novos modelos e permitir a combinação de modelos usando funções.

Outra vantagem é a proposta de uma metodologia para realizar a comparação entre modelos de RI sem a necessidade de realizar experimentos. Podemos comparar modelos através da formalização do conceito de equivalência entre modelos. A comparação entre modelos é importante devido às seguintes razões: para um melhor entendimento do relacionamento entre os modelos comparados,

para reutilização de código ou implementação de um modelo e para um melhor entendimento da semântica de similaridade. Neste trabalho, construímos modelos equivalentes a outros modelos, mas modelos existentes podem ser comparados usando nosso meta-modelo.

Como visto na literatura, a combinação de evidências pode melhorar a qualidade do *ranking*. Apresentamos dois estudos de caso usando a modelagem bayesiana e vetorial para combinação de múltiplas fontes de evidências, mas outras abordagens podem ser usadas. Outra contribuição de nosso trabalho é a proposta do uso da estrutura funcional como unificador de modelos de RI e a aplicação dessa estrutura para combinar fontes de evidências. A estrutura funcional pode ser usada para combinação de múltiplas evidências de vários modelos e ajudar no desenvolvimento de novos modelos e combiná-los, sendo a modelagem de evidências usando funções mais comum que a modelagem de evidências usando probabilidades.

Também apresentamos as idéias de desenvolvimento de uma ferramenta para comparação experimental entre modelos de RI baseada nos conceitos da estrutura funcional. Esta ferramenta é uma ferramenta genérica que auxilia pesquisadores e usuários de sistemas de recuperação de informação na avaliação da qualidade de modelos de RI.

7.2 Trabalhos Futuros

Os trabalhos futuros incluem os seguintes tópicos:

- A comparação entre outros modelos de RI e o estudo de novos modelos que sejam equivalentes aos modelos existentes, porém mais simples e de fácil implementação que os modelos existentes.
- O desenvolvimento de outros modelos para combinar múltiplas fontes de evidências usando outras semânticas de modelagem.
- O novo modelo baseado em distância pode ser usado para clusterização de documentos e no contexto de aplicações de alta dimensões.
- Uma interessante área de pesquisa seria estudar outras características dos modelos, por exemplo, definir algumas das propriedades que os modelos devem ter para que eles possuam maior

precisão ou revocação que outros (comparação relativa ou algébrica).

- A implementação da ferramenta proposta nesta dissertação.
- A realização de experimentos para verificar qual dos modelos dos estudos de caso apresentados neste trabalho possui melhor qualidade.

Referências Bibliográficas

[1] R. Baeza-Yates & B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] G. Bordogna & G. Pasi. Linguistic aggregation operators in fuzzy information retrieval. *International Journal of Intelligent Systems*, pg. 233–248, 1995.

[3] P. D. Bruza, F. Crestani, & M. Lalmas. Second workshop on logical and uncertainty models for information systems. In *Proceedings of DEXA*. IEEE Press, 2000.

[4] P. D. Bruza & T. W. C Huibers. Investigating aboutness axioms using information fields. In *In Proceedings of ACM SIGIR*, pg. 112–121, Dublin, Ireland, 1994.

[5] P. D. Bruza & Lalmas M. Logic based information retrieval: Is it really worth it? In *Proceedings of WIRUL 96, the SecondWorkshop on Information Retrieval, Uncertainty and Logic(Glasgow)*. 1996.

[6] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, & I. Silva. Local versus global link information in the web. *ACM Transactions On Information Systems*, 21(1):42–63, January 2003.

[7] C. Carrick & C. R. Watters. Automatic association of news items. *Information Processing Management*, 33(5):615–632, 1997.

[8] S. Chakrabarti, B. E. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. S. Tomkins, D. Gibson, & J. Kleinberg. Mining the link structure of the world wide web. *IEEE Computer*, 32(8):60–67, Agosto 1999.

[9] P. S. Chen. On inference rules of logic-based information retrieval systems. *Information Processing and Management*, pg. 43–59, 1994.

[10] Y. Chiaramella. About retrieval models and logic. *The Computer Journal*, pg. 233–241, 1992.

[11] F. Crestani & M. Lalmas. Logic and uncertainty in information retrieval. In *ESSIR*, pg. 179–206, 2000.

[12] F. Crestani & C. J. van Rijsbergen. Information retrieval by logical imaging. *Journal of Documentation*, 51:3–17, 1995.

[13] W. B. Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert: Intelligent Systems and Their Applications*, 8(2):8–12, 1993.

[14] W. B. Croft. Effective text retrieval based on combining evidence from the corpus and users. *IEEE Expert: Intelligent Systems and Their Applications*, 10(6):59–63, 1995.

[15] M. A. Pinheiro de Cristo, P. Calado, M. Silveira, I. Silva, R. Muntz, & B. A. Ribeiro-Neto. Bayesian belief networks for ir. *International Journal of Approximate Reasoning*, 34(2-3):163–179, 2003.

[16] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, & Marco Gori. Focused crawling using context graphs. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pg. 527–534, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[17] S. Dominich. Interaction information retrieval. *Journal of Documentation*, 50(3):197–212, 1994.

[18] S. Dominich. Formal foundation of information retrieval. In *Proceedings of the Workshop on Mathematical/Formal Methods in Information Retrieval at the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pg. 8–15, Athens, Greece, 2000.

[19] S. Dominich. A unified mathematical definition of classical information retrieval. *Jornal of the American Society for Information Science*, 51(7):614–624, 2000.

[20] S. Dominich. On applying formal grammar and languages, and deduction to information retrieval modelling. In *Proceedings of the ACM SIGIR MF/IR*, pg. 37–41, 2001.

[21] S. Dominich. Connectionist interaction information retrieval. *Inf. Process. Manage.*, 39(2):167–193, 2003.

[22] S. T. Dumais & R. Jin. Probabilistic combination of content and links. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pg. 402–403, New Orleans, Louisiana, USA, September 2001.

[23] B. W. Frakes & R. Baeza-Yates. *Information Retrieval and Data Structures*. Prentice Hall, 1992.

[24] N. Fuhr. Language models and uncertain inference in information retrieval. In *Proceedings of the Language Modeling and IR workshop*.

[25] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[26] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, pg. 33–44, Jan-Feb 1975.

[27] D. A. Grossman & O. Frieder. Information retrieval, algorithms and heuristics. Kluwer Academic Publishers, 1998.

[28] D. Harman. Overview of the third text retrieval conference. In *Proceedings of the Third Text Retrieval Conference - TREC-3*, Gaithersburg, Maryland, 1995. National Institute of Standards and Technology. NIST Special Publication 500-225.

[29] J. Horng & C. Yeh. Applying genetic algorithms to query optimization in document retrieval. *Inf. Process. Manage.*, 36(5):737–759, 2000.

[30] T. W. C. Huibers & P. D. Bruza. Situations: A general framework for studying Information Retrieval. In R. Leon, editor, *Information retrieval: New systems and current research, Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialists Group*, pg. 3–25. Taylor Graham, Drymen, Scotland, 1996.

[31] T. W. C. Huibers & P. D. Bruza. Situations: A general framework for studying Information Retrieval. In R. Leon, editor, *Information retrieval: New systems and current research, Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialists Group*, pg. 3–25. Taylor Graham, Drymen, Scotland, 1996.

[32] T. W. C. Huibers, M. Lalmas, & C. J. van Rijsbergen. Information retrieval and situation theory. *SIGIR Forum*, 30(1):11–25, 1996.

[33] E. M. Abinader Júnior. Combinação e avaliação de múltiplas fontes de evidências para recuperação de documento na web. Master's thesis, Universidade Federal do Amazonas, Instituto de Ciências Exatas, Amazonas, Manaus, 2004.

[34] H. Kang & K. Choi. Two-level document ranking using mutual information in natural language information retrieval. *Inf. Process. Manage.*, 33(3):289–306, 1997.

[35] L. Kaufman & P. J. Roussew. Finding groups in data - an introduction to cluster analysis. *Wiley-Science Publication John Wiley & Sons Inc.*, 1990.

[36] A. Kent, M. M. Berry, L. V. Luehrs Jr, & J. W. Perry. Machine literature searching VIII: Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93–101, 1955.

[37] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pg. 668–677, January 1998.

[38] J. Lafferty & C. Zhai. Probabilistic relevance models based on document and query generation. In *W. B. Croft and J. Lafferty, editors, Language Modeling and Information Retrieval*. Kluwer Academic Publishers, 2003.

[39] M. Lalmas & P. D. Bruza. The use of logic in information retrieval modeling. *Knowledge Engineering Review. In press.*, 13(3):263–295, 1998.

[40] J. H. Lee, W. Y. K., M. H. Kim, & Y. J. Lee. On the evaluation of boolean operators in the extended boolean retrieval framework. In *SIGIR 93: Proceedings of the 16th annual international*

ACM SIGIR conference on Research and development in information retrieval, pg. 291–297, New York, NY, USA, 1993. ACM Press.

[41] R. Lempel & S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160, April 2001.

[42] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[43] J. Y. Nie. *Un Modèle de Logique Générale pour les Systemes de Recherche d'Informations. Application au Prototype RIME*. PhD thesis, Université Joseph Fourrier, Grenoble, France, 1990.

[44] L. Page, S. Brin, R. Motwani, & Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Technical report, Stanford Digital Library Technologies Project, 1998.

[45] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.

[46] A. Montejo Ráez. Formal models for ir: a review and a proposal for keyword assignment. In *Workshop on Mathematical/Formal Methods in Information Retrieval*. ACM-SIGIR, 2003.

[47] B. Ribeiro (Ribeiro-Neto) & R. Muntz. A belief network model for ir. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pg. 253–260, Zurich, Switzerland, 1996.

[48] S. E. Robertson. Salton award lecture: On theoretical argument in information retrieval. *SIGIR Forum*, 34(1):1–10, 2000.

[49] R. Rousseau. Extended boolean retrieval: a heuristic approach? In *SIGIR 90: Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pg. 495–508, New York, NY, USA, 1990. ACM Press.

[50] S. Russell & P Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1995.

[51] G. Salton. Automatic indexing using bibliographic citations. *Journal of Documentation*, 27(2):98–110, 1971.

[52] G. Salton & C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[53] G. Salton & M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.

[54] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, & N. Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pg. 96–103, Athens, Greece, July 2000.

[55] I. Silva, J. N. Souza, R. Moura, & Ribeiro-Neto. Informação de Links no Modelo Vetorial Usando uma Estrutura Funcional. In *Anais do 18th Simpósio Brasileiro de Banco de Dados*, pg. 170–184, Manaus, AM, Brasil, 2003.

[56] I. Silva, J. N. Souza, & L. Oliveira. Comparison of Representations of Multiple Evidence Using a Functional Framework for IR. In M. P. Consens & G. Navarro, editors, *Proceedings of the 12th International Symposium on String Processing and Information Retrieval (SPIRE)*, Published as Lecture Notes in Computer Science 3772, pg. 283–294. Springer, November 2005.

[57] C. Silverstein, M. R. Henzinger, H. Marais, & M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[58] D. Song & P. Bruza. Discovering information flow using a high dimensional conceptual space. In *Proceedings of the 24th ACM SIGIR Conference*, pg. 9–12. ACM Press, 2001.

[59] D. W. Song, K. F. Wong, P. D. Bruza, & Cheng C. H. Towards a commonsense aboutness theory for information retrieval modeling. In *In Proceedings of the FourthWorld Multiconference on*

Systemics, Cybernetics and Informatics (SCI 2000), pg. 23–26, Orlando, Florida (USA), July 2000.

[60] A. Spink, DietmarWolfram, B. J. Jansen, & T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, February 2001.

[61] A. Spink, B. J. Jansen, Dietmar Wolfram, & T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109, April 2002.

[62] L. T. Su. Value of search results as a whole as the best single measure of information retrieval performance. *Inf. Process. Manage.*, 34(5):557–579, 1998.

[63] T. Tsikrika T. Roelleke and & G. Kazai. A general matrix framework for modelling information retrieval. *Journal on Information Processing & Management (IP&M), Special Issue on Theory in Information Retrieval, to appear*, 2005.

[64] H. Turtle & W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[65] C. J. van Rijsbergen. *Information Retrieval*. Butterwords, 1979.

[66] C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6), 1986.

[67] J. Verhoeff, W. Goffman, & Jack Belzer. Inefficiency of the use of boolean functions for information retrieval systems. *Commun. ACM*, 4(12):557–558, 1961.

[68] Jr. W. D. Morris & J. Lawrence. Geometric properties of hidden Minkowski matrices. *SIAM Journal on Matrix Analysis and Applications*, 10(2):229–232, 1989.

[69] T. Westerveld, W. Kraaij, & D. Hiemstra. Retrieving Web pages using content, links, URLs and anchors. In *The Tenth Text Retrieval Conference (TREC-2001)*, pg. 663–672, Gaithersburg, Maryland, USA, November 2001.

[70] K. F. Wong, D. Song, P. Bruza, & C. H. Cheng. Application of aboutness to functional benchmarking in information retrieval. *ACM Trans. Inf. Syst.*, 19(4):337–370, 2001.

[71] J. Zobel & A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.