CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Elias Luiz da Silva Júnior

ANÁLISE DE DESEMPENHO DE UMA IMPLEMENTAÇÃO DE UNIDADE DE MEMORIZAÇÃO DE TRAÇOS DINÂMICOS EM FPGA

Timóteo

2016

Elias Luiz da Silva Júnior

ANÁLISE DE DESEMPENHO DE UMA IMPLEMENTAÇÃO DE UNIDADE DE MEMORIZAÇÃO DE TRAÇOS DINÂMICOS EM FPGA

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Bruno Rodrigues Silva

Timóteo

2016

.

.

.

Dedico a
algumas pessoas.

Agradecimentos

Agradeço aos pais no primeiro parágrafo?

E à namorada ou ao namorado no segundo?

Orientador no terceiro?

Esta página de agradecimentos vive dando problemas.

“Os grandes navegadores
devem sua reputação aos temporais e tempestades”.
Epicuro

Resumo

Um único parágrafo que sintetize todo o meu trabalho. Organização de informação usando classificação facetada é útil para melhorar a indexação de documentos e a construção de sistemas de recuperação de informação corporativa. Essa hipótese baseia-se na evidência de facetas comuns a documentos de diferentes empresas e na flexibilidade da organização facetada. Entretanto, a classificação e indexação automáticas de um grande volume de documentos representam importantes obstáculos e nossa principal motivação. A pesquisa é descritiva, aplicada e experimental e tenta responder sobre a existência de características comuns a documentos do domínio corporativo e a possibilidade de indexação facetada automática. Duas coleções são usadas para avaliação, uma pública e outra particular. Os termos usados por autores de documentos foram obtidos através de documentos e expressões de busca. Foi empreendida uma análise preliminar do domínio corporativo, pela qual foram descobertas 12 categorias comuns e facetas úteis para o contexto de cada coleção de avaliação. A distribuição de assuntos em categorias apresentou alta correlação positiva usando o coeficiente de correlação de Spearman. Dez expressões de busca de usuários foram avaliadas no contexto da coleção particular e validaram as 12 categorias comuns. A avaliação empírica da trilha *Enterprise* da *Text Retrieval Conference* foi executada e os métodos de indexação, classificação e recuperação automáticos de informação facetada melhoraram a eficiência da recuperação sem fazer uso de serviços externos, como Wikipedia e metabuscadores, e sem fazer uso de estruturas hipertextuais presentes nos documentos da amostra. A avaliação empírica utilizou-se principalmente das características espaciais, temporais, de documento e de pessoal. A técnica de análise facetada mostrou-se promissora para os métodos de análise e comparação de coleções corporativas sem que dados puros sejam expostos a terceiros. A tese aponta direções de pesquisa para o uso dos métodos em outras coleções, para aperfeiçoamentos da organização da informação facetada, e para novas aplicações dos métodos também em outros domínios.

Palavras-chave: análise de domínio, análise facetada, recuperação de informação, informação corporativa.

Abstract

We hypothesise that information organisation based on faceted classification is useful to improve enterprise information retrieval systems. The existence of similar facets in documents from different companies and the known adaptability of facet organisation strengthen this hypothesis. We refer this work to the automated classification and indexing on large amounts of text files. This work is descriptive, applied, and experimental. It aimed to expose the main characteristics of the enterprise information, proposing a tentative generalisation to the enterprise domain and presenting some facets we can use to organise it and to support better information retrieval. It applied facet analysis to two enterprise collections and evaluated the resulting faceted classification. Terms were selected from documents and queries. We found twelve common categories and the distribution of document subjects across the categories presents strong positive correlation by the Spearman's rank correlation. Then, we obtained ten user queries and we adopted them to validate the found categories. We also used the Enterprise track of Text Retrieval Conference and its previous results as a Cranfield-like evaluation. The automated prototype used spatial, temporal, document and social characteristics. Thus, our empirical evaluation improved the information retrieval with no external dependency like Wikipedia or metasearch engines. The facet analysis was useful for comparing the companies with no desire to expose their information. The method can guide and stimulate future work and other companies can become more willing to take part in a research study.

Keywords: domain analysis, facet analysis, information retrieval, enterprise information.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A Lei de Moore possui algumas variações quanto ao seu enunciado, porém todas afirmam que a capacidade computacional dos processadores cresceria exponencialmente devido aos avanços na tecnologia. Por 50 anos essa previsão se manteve consistente com os produtos lançados no mercado, como descrito em. Porém, limitações físicas na criação de circuitos integrados ameaçam o continuidade dessa evolução. (**??**)

Mas com o crescente aumento da demanda por computação é necessário que os projetistas encontrem maneiras de aperfeiçoar ainda mais o funcionamento das unidades de processamento. Uma solução que vem sendo utilizada é acoplar vários processadores para funcionar em paralelo, porém isso aumenta a complexidade de projetos tanto a nível de hardware como de software, além de amplificar o consumo energético do sistema.

O grande desafio da arquitetura de computadores é buscar soluções eficientes, conciliando fatores como desempenho do sistema, consumo de energia, custo de produção e tamanho e complexidade do produto final. Em muitas situações, esses fatores concorrem entre si, levando o projetista a ter de tomar decisões sobre qual abordagem será escolhida para solucionar determinado problema.

O que ocorre então é a criação de sistemas especialistas para determinadas funções, enquanto outros projetos mais gerais lidam com uma gama mais diversa de aplicações. Em ambos os casos, projetistas consideram qual problema buscam resolver para criar a solução mais adequada dentro das restrições.

Como exemplo podemos comparar as diferentes abordagens assumidas ao projetar um *system-on-chip* para aplicação em um sistema embarcado e na criação de uma unidade de processamento gráfico. Enquanto sistemas embarcados prezam por tamanho reduzido e baixo consumo de energia, unidades gráficas têm como prioridade a velocidade para cálculos de ponto flutuante, sendo otimizadas para executar instruções simples a diversos dados de entrada simultaneamente. (**??**)

Assim, é importante conhecer e desenvolver técnicas que possam tornar os projetos mais eficientes. Desenvolver para que o custo-benefício do produto seja melhorado independentemente de avanços na tecnologia de produção, mas sim por um design melhor elaborado. Conhecer para que seja possível ponderar como e quais técnicas aplicar para que o objetivo final possa ser atingido de maneira ótima, com um máximo de desempenho e mínimo de recursos despendidos.

1.1 Justificativa

1.2 Problema

As seguintes questões constituem o problema desta pesquisa: A informação corporativa possui características que potencialmente sejam comuns a todo o domínio corporativo? No contexto de uso de sistemas de recuperação de informação corporativa, as expressões de busca de usuários apresentam as mesmas características que estão presentes nos documentos? A organização facetada da informação corporativa contribui para o aumento do desempenho geral do sistema automático de recuperação de informação?

As características da informação, em documentos e expressões de busca, referem-se a atributos que devem ser reconhecidos a partir do domínio corporativo. Adicionalmente, a contribuição esperada da organização facetada da informação está associada à sua capacidade de representar entidades por meio da totalidade ou de um subconjunto de suas características, suportando a recuperação de entidades e consequentemente a recuperação de documentos onde as entidades estão contidas. Finalmente, o desempenho do sistema de recuperação de informação refere-se à sua capacidade atender às necessidades informacionais de seus usuários, algo que pode ser parcialmente medido através de métricas discutidas na literatura.

1.3 Objetivos

Para responder ao problema proposto, o objetivo geral deste trabalho é propor um conjunto de características da informação corporativa que favoreça a organização e a recuperação da informação.

Também, objetivam-se mais especificamente:

1. Propor um conjunto de facetas que seja útil na organização automática da informação corporativa e na interoperabilidade entre diferentes repositórios de informação da mesma empresa ou de diferentes empresas;
2. Identificar as facetas pelas quais os usuários de informação especificam sua necessidade de informação no contexto de trabalho;
3. Avaliar as implicações da organização facetada da informação corporativa no desempenho dos sistemas automáticos de recuperação de informação.

Para responder as questões propostas e aos objetivos apresentados, pretende-se empreender uma análise preliminar do domínio corporativo a partir de dois exemplares do domínio, avaliar como os usuários informacionais mobilizam as facetas mais comuns do domínio, e implementar e avaliar um protótipo de sistema de recuperação de informação corporativa. Os procedimentos metodológicos seguidos são apresentados no capítulo 2.

1.4 Estrutura da tese

Esta tese está estruturada em seis capítulos, ordenados pelo momento em que foram concluídos dentro do ciclo de vida desta pesquisa, e anexos, a saber:

- O capítulo 2 apresenta os procedimentos metodológicos através dos quais este trabalho se desenvolve, o que inclui a formação das coleções de documentos para experimentação e avaliação, as etapas para o projeto do protótipo funcional, bem como as estratégias de validação.
- As bases teóricas são apresentadas em seguida, no capítulo 3, o que inclui marcos conceituais importantes para o contexto de sistemas de recuperação de informação corporativa e para a organização do conhecimento e da informação em estruturas facetadas.
- Uma análise preliminar de domínio é descrita no capítulo 4. Duas coleções de documentos são investigadas para identificar as facetas mais comuns que possam constituir um conjunto mínimo de facetas para representar entidades no domínio corporativo.
- No capítulo 5 é detalhada uma avaliação da coleção pública e é obtido um conjunto de expressões de busca a partir de seus usuários sobre uma das coleções estudadas. São então avaliadas a utilidade e a eficiência de um modelo de recuperação baseado em facetas do domínio corporativo.
- Finalmente, no capítulo 6 são apresentadas as conclusões e considerações finais, principais contribuições, limitações e indicadas algumas direções para trabalhos futuros.
- A tese inclui parte dos seus resultados e produtos em anexos, tendo em vista que sua extensão poderia comprometer a legibilidade do texto e a compreensão do leitor. As coleções estudadas não são disponibilizadas entre os anexos pois isso violaria alguns direitos de propriedade intelectual dos seus autores. A tese aponta outros meios de obter uma cópia das referidas coleções.

2 Procedimentos metodológicos

“Um bom começo é a metade”.
Aristóteles

A presente pesquisa é descritiva e exploratória do ponto de vista dos objetivos; aplicada do ponto de vista de sua natureza; qualitativa e quantitativa quanto à abordagem ao problema; e pode ser classificada como pesquisa experimental na perspectiva dos procedimentos técnicos, embora tenha mobilizado diferentes procedimentos técnicos em diferentes partes, que merecem classificação diferenciada.

Os procedimentos metodológicos são organizados nas seguintes etapas:

1. reunir e estudar os principais trabalhos de organização de informação, especialmente aqueles orientados a coleções corporativas – parte da pesquisa que pode ser classificada como bibliográfica e tornou-se fundamental para reconhecer características do domínio corporativo que já foram explicitadas por outros autores, de diversas áreas do conhecimento;

2. reunir dois exemplares do domínio corporativo, sendo que um deles é uma coleção pública, altamente reconhecida na literatura; e o outro é uma coleção particular criada especificamente para o trabalho desta tese, resultado da coleta e reunião de uma massa de dados convencional de uma empresa, com diferentes tipos de documentos, idiomas e gêneros linguísticos – parte da pesquisa que pode ser classificada como pesquisa documental;

3. propor um conjunto mínimo de facetas que represente a informação corporativa das duas empresas investigadas – parte da pesquisa que pode ser classificada como pesquisa documental;

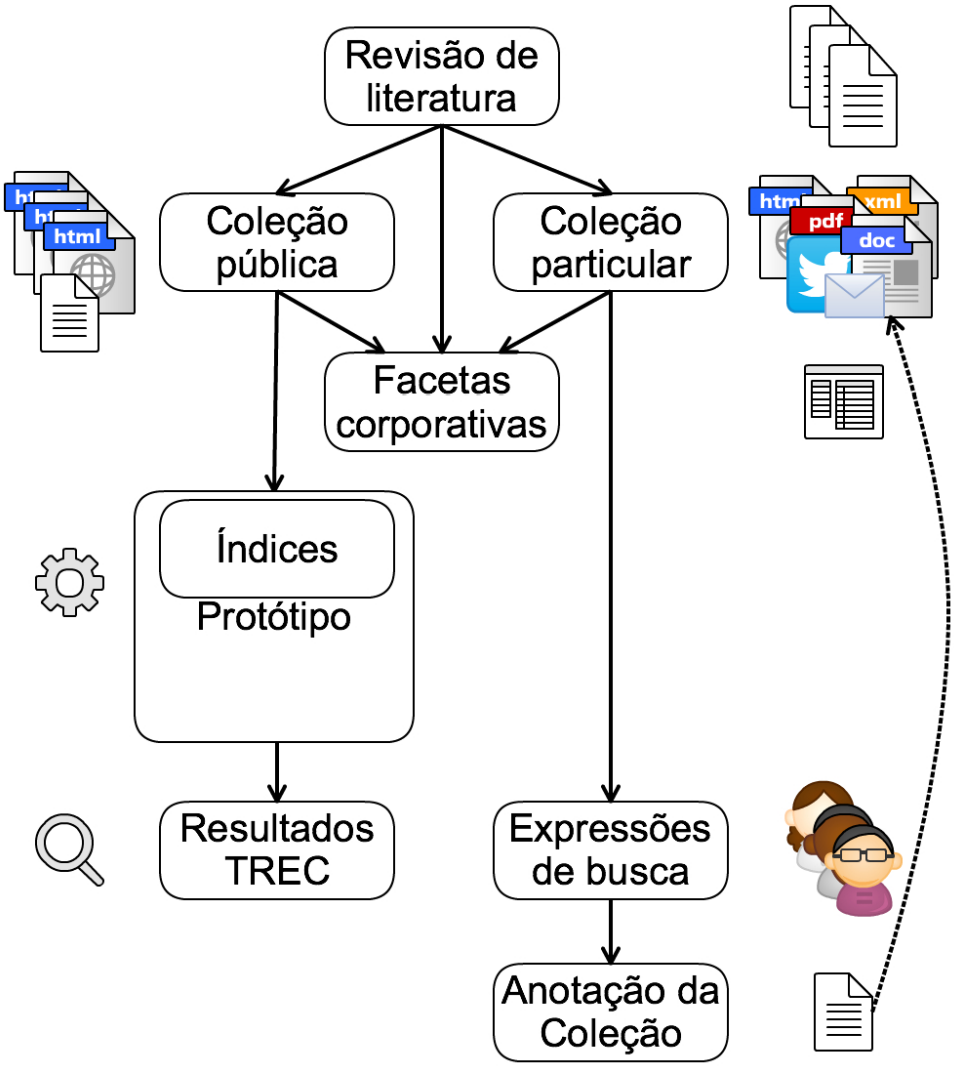
4. projetar e implementar um arcabouço de *software* que indexe a massa de dados de documentos corporativos usando estruturas de dados específicas – parte da pesquisa que pode ser classificada como estudo de caso;

5. avaliar os resultados usando as expressões de busca dos usuários da informação para recuperar documentos da coleção particular – parte da pesquisa que pode ser classificada como levantamento;

6. avaliar o desempenho da organização facetada utilizando-se do método de avaliação de Cranfield, observando os resultados da trilha *Enterprise* da *Text Retrieval Conference* (TREC) – parte da pesquisa que pode ser classificada como pesquisa experimental;

7. preparar a coleção particular para que sirva como proposta para avaliação de técnicas de recuperação de informação corporativa em conjuntos heterogêneos de documentos

Figura 1 – Procedimentos metodológicos



Fonte: elaborada pelo autor

quanto ao tipo e gênero textual, em língua portuguesa – parte da pesquisa que pode ser classificada como pesquisa participante.

As diferentes etapas são descritas mais detalhadamente nas próximas seções do presente capítulo e sintetizadas na figura 1.

2.1 Revisão da literatura

A revisão de literatura empregada neste trabalho resulta em dois conjuntos distintos de trabalhos relacionados: i) a revisão do estado da arte em análise de domínio e recuperação de informação; e ii) o levantamento bibliográfico dos principais resultados de trabalhos anteriores que explicitaram características do domínio corporativo, com origem em diversas áreas do conhecimento.

O estado da arte e da técnica em análise de domínio e recuperação da informação é apresentado no capítulo 3. O contexto corporativo foi adotado para limitar o escopo deste trabalho e torná-lo viável para estudos com usuários e para classificação intelectual de documentos. Com isso, trabalhos relacionados mais especificamente à informação corporativa são especialmente de interesse. Na literatura sobre informação corporativa, também se busca um conjunto de facetas para a informação corporativa, como se observa na figura 1. Características da informação corporativa naturalmente permeiam os produtos tecnológicos de recuperação de informação corporativa e são discutidas em trabalhos sobre o tema nas áreas de Ciência da Computação e Biblioteconomia e Ciência da Informação. Entretanto, a fragmentação desse conhecimento torna difícil produzir um mapa das características do domínio corporativo. Adicionalmente, como os trabalhos tendem a assumir uma perspectiva mais pragmática, as características corporativas já conhecidas tendem a ser úteis e restritas apenas ao seu contexto original de estudo.

O segundo produto da revisão engloba um conjunto com os principais métodos automáticos e semiautomáticos de classificação, indexação e *ranking* de informação, bem como as métricas usadas para avaliação experimental do desempenho dos métodos. Devidamente estudados, devem servir de base de comparação para projetar métodos mais adequados para organizar e recuperar informação corporativa. O segundo produto é apresentado na seção **??**, precisamente onde é adotado para validar os resultados deste trabalho sobre a coleção pública.

2.2 Coleções de documentos

Neste trabalho, duas coleções de documentos corporativos são adotadas para a análise de domínio. A primeira coleção refere-se à coleção de referência usada na trilha *Enterprise* da *Text Retrieval Conference* (TREC) até o ano de 2008. Trata-se de uma coleção de 370.715 páginas *Web* públicas da *Commonwealth Scientific and Industrial Research Organisation* (CSIRO). A segunda coleção refere-se a um conjunto mais amplo de documentos, o que inclui atas, relatórios, memorandos, e-mails e páginas *Web* públicas.

A coleção de referência da *Text Retrieval Conference* é útil por facilitar a comparação entre experimentos empíricos relatados na literatura e experimentos empreendidos nesta tese. Porém, seu uso tem sido criticado por contar exclusivamente com páginas *Web* pública da CSIRO, criando condições que em nada se parecem com aquelas presentes em um ambiente real de busca corporativa.

Na tentativa de reduzir as limitações da coleção de referência, é adotada uma segunda coleção de documentos. Trata-se de um conjunto de documentos de uma empresa pública brasileira, com uma diversidade maior de tipos de documentos que aquela encontrada na coleção de referência. Ela se caracteriza como um ambiente mais próximo da realidade de uma empresa e das necessidades de informação de um usuário corporativo real, porém não pode ser vista como uma substituta da coleção de referência da *Text Retrieval Conference* e nem mesmo como uma coleção perfeita para todo e qualquer objetivo. Nesta tese, a segunda coleção é denominada como coleção particular. No entanto, a coleção também se tornará

publicamente disponível no ano de 2015.

Documentos da coleção pública são classificados para diferentes tarefas de busca. As tarefas de busca são aquelas mais comumente realizadas por usuários de informação reais, responsáveis pela classificação intelectual dos documentos para a avaliação, sendo que os usuários não são conhecidos. Ao contrário, para estudos sobre a coleção particular está disponível uma amostra de usuários reais.

A presença de duas coleções provoca impactos em vários procedimentos metodológicos, como a figura 1 ilustra, requerendo dois experimentos e avaliações diferentes. Enquanto a coleção particular faz uso de seus usuários para validar os resultados deste trabalho, a coleção pública conta com resultados prévios obtidos em duas edições da *Text Retrieval Conference*, em 2007 e 2008.

Finalmente, a coleção particular tem sido usada por seus usuários para a realização de trabalho e para a tomada de decisão nos últimos anos. Portanto, partindo do pressuposto que a coleção particular seja uma coleção corporativa válida, pretende-se validar também a coleção pública como uma coleção corporativa ao demonstrar a compatibilidade de ambas.

2.3 Definição de um conjunto mínimo de facetas corporativas

O estabelecimento de um conjunto mínimo de facetas, que possa atender adequadamente as necessidades de diferentes empresas, requer uma análise de domínio. A partir das duas coleções citadas na seção anterior, a análise de domínio adota as técnicas de análise de assuntos e de análise facetada. Ambas as técnicas servem ao propósito de descobrir características dos documentos corporativos e propor um conjunto de facetas comuns a ambas as coleções.

A análise facetada permite ampliar facilmente esse conjunto de características, acomodando-as em esquemas classificatórios hospitaleiros. Isso é útil para permitir que um esquema classificatório mais generalista seja personalizado para uma empresa específica, ou para uma unidade organizacional específica. Por outro lado, um conjunto genérico e realmente expressivo de características é útil para a interoperabilidade entre sistemas de informação, para o intercâmbio de informação corporativa, e principalmente para o projeto mais eficiente de sistemas de recuperação de informação corporativa.

Os procedimentos metodológicos do processo de análise de domínio, através da análise de assunto e da análise facetada, são descritos no capítulo 4.

2.4 Prototipação de um sistema de recuperação de informação corporativa

A prototipação baseada em *software* é descrita na figura 1 apenas como protótipo. O protótipo de um sistema de recuperação de informação, implementado em linguagem Java e usando a biblioteca Lucene, executa as funções de indexar e recuperar documentos apenas da

coleção pública para realizar um experimento empírico. O protótipo é documentado no capítulo 5.

A metodologia de prototipação se baseia no trabalho de Anastácio (2009), pela adoção de mecanismos comuns de coleta, classificação, indexação, busca e recuperação de documentos e sua adaptação para necessidades especiais. No contexto desta tese, os requisitos do protótipo são: organização facetada, tratamento espaço-temporal, e reconhecimento de entidades sociais, espaciais e temporais. Pela implementação desses requisitos, um sistema de recuperação de informação comum torna-se um sistema de recuperação de informação corporativa e facetada.

Os requisitos não-funcionais do protótipo incluem o tratamento geográfico através da indexação espacial e da implementação de um *gazetteer*; o tratamento temporal através da indexação de indicadores de tempo; e a federação dos repositórios corporativos através da extração de texto dos documentos coletados e a geração de um índice centralizado. Os requisitos funcionais do protótipo corporativo e facetado incluem a indexação, a interface de busca com o usuário, a busca em lote para avaliação de várias expressões de busca em conjunto, e a recuperação de informação.

O controle de acesso a documentos, um importante requisito não-funcional dos sistemas de recuperação de informação corporativa, não é considerado pela natureza da coleção de documentos da CSIRO, constituída apenas por dados acessíveis a todos os usuários. Os efeitos dessa omissão não afetam o cenário de avaliação proposto pela trilha *Enterprise* da *Text Retrieval Conference*.

Essa avaliação baseada em prototipação constitui um importante método de validação de resultados desta tese. Porém, sua utilidade é limitada apenas a uma das duas coleções de documentos adotadas, como se observa na figura 1. A validação de resultados da coleção particular se dá através dos usuários de informação e não beneficia-se diretamente desse método experimental.

2.5 Avaliação e validação

Dois cenários de avaliação são adotados e ilustrados na figura 1. O primeiro cenário de avaliação refere-se à avaliação experimental da trilha *Enterprise* da *Text Retrieval Conference*, pela comparação dos resultados desta tese com aqueles de outros trabalhos disponíveis na literatura. Os principais resultados experimentais foram publicados no ano de 2008, quando a trilha foi extinta. O segundo, por sua vez, refere-se exclusivamente à coleção particular e baseia-se em seus usuários. Esse cenário de avaliação usa as expressões de busca elaboradas pelos usuários de informação para avaliar se as características da informação corporativa são reconhecidas por seus usuários.

Os métodos de avaliação e validação são detalhados no capítulo 5, juntamente com os resultados da avaliação, sua análise e discussões.

2.6 Anotação da coleção particular

A coleção pública de documentos, ou coleção de referência, conta com anotações feitas por especialistas. Tais anotações referem-se a caracterização da coleção, listagem de documentos e pessoas relevantes, e de buscas mais frequentes. A coleção particular, usada primeiramente nesta tese, necessita desse tipo de caracterização para ser usada em futuros trabalhos sobre organização da informação corporativa.

Como demonstrado na figura 1, a atividade de anotação dessa coleção é realizada a partir da documentação do experimento sobre a coleção particular, constituindo o último produto desta tese. Após a disponibilização de uma réplica da coleção particular, por seus proprietários, a coleção e a sua anotação garantem que os resultados desta tese podem ser repetidos e validados em trabalhos futuros.

3 Fundamentos históricos, teóricos e metodológicos

"O período de maior ganho em conhecimento e experiência é o período mais difícil da vida".
Dalai Lama

O principal objetivo deste capítulo é apresentar os fundamentos históricos, teóricos e metodológicos sobre os quais esta pesquisa é executada, além de mapear os principais e mais recentes trabalhos em análise de domínio e organização da informação corporativa. As seções seguintes reúnem os principais e mais recentes trabalhos que versam sobre análise de domínio e organização da informação que impliquem direta ou indiretamente em processos de classificação, indexação e recuperação de informação corporativa.

3.1 Análise de domínio

O objetivo desta tese é a caracterização do domínio corporativo visando favorecer a atividade de recuperação da informação corporativa. Ou seja, o objetivo é a descoberta das características que o domínio corporativo possui e apresenta, implícita e explicitamente, aos membros da sua comunidade; e a exploração dessas características para o aperfeiçoamento de sistemas automáticos de recuperação de informação corporativa.

A formalização das características do domínio é o produto de um processo denominado análise de domínio (ALVARENGA; DIAS, 2012), pelo qual as características tornam-se explícitas através da observação e da interpretação do domínio em seu contexto de produção e uso. Um domínio, porém, é uma comunidade discursiva imersa em uma história, uma cultura, uma janela de tempo e um espaço. Portanto, o domínio refere-se a uma entidade intangível e a análise de domínio precisa partir desse pressuposto (HJØRLAND; ALBRECHTSEN, 1995).

Para enfrentar a intangibilidade do domínio, a análise de domínio dá-se através de abordagens que criam a impressão de que o domínio seja modular, sendo que a compreensão do todo dar-se-ia pela análise e compreensão dos módulos que o compõem. Em linhas gerais, "análise é feita com base nas informações oriundas das comunidades discursivas, a partir da sua linguagem e de suas condições culturais e históricas" (DIAS; ALVARENGA, 2011, p. 182). Então, o processo da análise de domínio começaria pela definição da amostra ou dos exemplares do domínio que tornar-se-ão objeto de observação e análise.

Entretanto, abordagens bem-sucedidas na análise de um domínio não devem ser consideradas em outros, como se todos os domínios fossem similares (HJØRLAND, 2002). Alvarenga e Dias (2012) explicam usando o domínio corporativo como exemplo:

> a análise de domínio compreende o levantamento e estruturação dos entes que

> compreendem a realidade ôntica da empresa, como ser organizacional. Normalmente o trabalho de rastreamento de entes de um domínio acadêmico ou discursivo, na ciência da informação, é feito via literatura, tendo como suporte a garantia da literatura; neste caso os focos são campos de conhecimento mais ou menos sedimentados, não sendo, entretanto o caso dos seres organizacionais. Para o exercício de suas funções relativas às áreas meio e áreas finalísticas, as empresas têm como meta a identificação e descrição dos entes que a compõem, de suas essências, acidentes e processos que deles decorrem. Nesse desafio as entidades da realidade empresarial nem sempre se encontram devidamente identificadas e caracterizadas, na literatura técnico-científica publicada, mas podem estar presentes em documentos, administrativos, políticos, legislativos, etc. (ALVARENGA; DIAS, 2012)

Assim, definidos os exemplares do domínio, se deve escolher as abordagens de análise de domínio que melhor correspondem aos objetivos da análise, adequados ao domínio estudado e compatíveis com a área do conhecimento e com a área de formação dos analistas do domínio. Como exemplo, os "processos arquivísticos podem ser vistos como verdadeiras abordagens de análise de um domínio" (ALVARENGA; DIAS, 2012). Também, Hjørland (2002) enumera outras 11 abordagens utilizadas em diferentes áreas do conhecimento que são apropriadas para a análise de domínio. Dentre elas, algumas são de interesse para este trabalho, como métodos de classificação e construção de tesauros; métodos de indexação e recuperação de informação auxiliadas por computador; estudos bibliométricos sobre coleções de documentos corporativos; estudos de documentos e gêneros; estudos terminológicos, sobre linguagens documentárias e de discurso; e estudos sobre semântica em bancos de dados.

As abordagens citadas anteriormente podem ser usadas, isoladamente ou conjuntamente, no processo de análise de domínio. Os produtos de cada abordagem citada são conhecidos pela área da Biblioteconomia e Ciência da Informação: tesauros; *gazetteers*; modelos de recuperação de informação; linguagens de indexação; classificações; metadados; padrões de intercâmbio de dados e outros (LYKKE-NIELSEN, 2011). Para a análise de domínio, esses são apenas subprodutos (ALBRECHTSEN, 1993). No exemplo específico do domínio corporativo, "a análise de domínio, tal como preconizada na ciência da informação, prioritariamente se volta ao planejamento e desenvolvimento de sistemas de recuperação de informações empresariais, mas sabe-se que ela se encontra dentre os interesses de outros campos profissionais e de pesquisa" (ALVARENGA; DIAS, 2012). Ou seja, o produto da análise de domínio é um modelo de alto nível que favorece o projeto de sistemas de informação e serviços de informação, além de outros produtos de interesse em outros campos.

O produto principal da análise de domínio, um modelo de alto nível que também pode ser denominado como 'análise de domínio' ou 'uma análise de domínio', possui valor limitado pela representatividade dos exemplares usados para os estudos; profundidade pela qual o domínio foi observado; expressividade das abordagens usadas para empreender o processo de análise; e pela utilidade dos subprodutos colocados à disposição dos profissionais da informação que trabalham dentro do domínio. Portanto, a análise de domínio não produz um produto final, mas um modelo que deve ser aperfeiçoado na medida em que o domínio é descoberto, e que deve ser atualizado na medida em que o domínio se desenvolve.

3.1.1 Fontes para o processo de análise de domínio

Diversas fontes constituem matéria-prima potencial para o processo de análise de domínio, as quais incluem literatura técnica-científica do domínio, exemplares documentais, entrevistas com profissionais do domínio, dentre outras. A escolha de uma ou várias dessas fontes refletem a disponibilidade e a representatividade das fontes. Especificamente no contexto do domínio corporativo, são encontrados baixa cobertura da literatura, baixa disposição para participar de estudos que envolvam informação corporativa, alto sigilo profissional e grande variedade de produtos de informação que dificulta seu uso em análise de domínio. Os documentos e os repositórios corporativos de várias empresas parecem ser o ponto de partida mais apropriado para empreender a análise de domínio, embora ainda sejam difíceis de obter.

São vários os desafios de gerenciar repositórios de informação e recuperar documentos a partir desses repositórios. Um desses desafios refere-se a descrição dos documentos para subsidiar a construção de ferramentas de busca que suportem a pesquisa, a navegação, a recuperação e até mesmo a ordenação adequada de resultados (BROUGHTON, 2006). Tal descrição certamente depende de uma análise rigorosa do conteúdo do documento que facilite sua representação, retratando o significado, os atributos e relacionamento do documento com outros objetos (BRÄSCHER; CAFÉ, 2008). Deve-se descrever também aquilo que não está explicitamente presente no texto, como é o caso de conceitos e as relações mais diversas de espaço, de tempo ou de papeis (ALVARENGA; DIAS, 2012). Esse significado deve ser anotado utilizando-se de símbolos não necessariamente encontrados no texto. Esses símbolos permitem a representação da informação através de uma notação própria a qual pode ser chamada de metalinguagem (GARDIN, 1973), linguagem documentária (TÁLAMO, 2001) ou simplesmente linguagem de indexação (FUJITA, 2004).

As linguagens de indexação representam a informação por alguma estrutura que reflete os objetivos do sistema de organização de conhecimento no qual é aplicada. Hjørland (2012) apresenta uma taxonomia de sistemas de organização de conhecimento com diferentes estruturas, como listas de termos (glossários, dicionários, *gazetteers*), classificações e categorias (cabeçalhos de assunto, taxonomias, esquemas de classificação), e esquemas de categorização (listas de relacionamento, tesauros, redes semânticas e ontologias).

Há alguma compatibilidade entre as diferentes estruturas e portanto a transformação de uma estrutura em outra é natural. Porém, não é nada trivial traduzir o significado de uma estrutura para outra, uma vez que cada sistema controlado é construído com o propósito de representar um domínio a partir das necessidades particulares de seus usuários. Buscas federadas ou integradas tornam-se então pouco eficientes, quando não se mostram impossíveis (HALEVY et al., 2005; HJØRLAND, 2012). Ao tentar localizar e compreender informação de múltiplos repositórios em organizações, por exemplo, Halevy et al. (2005) afirmam que frequentemente encontram-se situações em que a informação necessária parece não ser capturada em qualquer repositório, ou é necessário um esforço significativo para compreender relações semânticas entre os repositórios e só então integrá-los.

Apesar da existência de métodos para agregar significado a dados e garantir a ponte semântica entre diferentes modelos de um mesmo domínio, a manutenção dos modelos e da

compatibilidade semântica no tempo mostra-se ainda mais desafiadora (HALEVY et al., 2005). Essa incompatibilidade semântica entre sistemas de informação não parece nada simples de resolver, uma vez que há incompatibilidade até mesmo entre os modelos de indexação e as ideias, ou modelos mentais, que os usuários têm da informação. Quando os modelos dos usuários e dos indexadores não coincidem sua utilidade fica comprometida e leva o usuário inevitavelmente ao sentimento de frustração e ao abandono do sistema de informação (HJØRLAND; ALBRECHTSEN, 1995; SOLOMON, 2002).

Para a área de Biblioteconomia e Ciência da Informação, a função da organização de informação inclui processos que partem da produção, seleção e coleção de objetos informacionais, passam pelo trabalho de representação, atribuição de entidades a classes de um sistema de classificação, recuperação e acesso; e chegam ao uso e implicações do uso da informação pela sociedade (SOLOMON, 2002). Porém, o escopo desta pesquisa está restrito aos processos de representação e classificação visando a recuperação de informação. Mesmo assim, há muitos outros processos importantes que vão além da recuperação de objetos relevantes. Isso reflete uma mudança de paradigma que trata do conjunto de objetos mais pertinente, ao invés de mais relevante, para suportar análises e tomada de decisão por pessoas no contexto de seu trabalho (SOLOMON, 2002). Esse requisito é especialmente sério quando visto dentro dos sistemas de recuperação de informação corporativa, dado seu caráter mais pragmático.

O primeiro processo de interesse, de representação da informação, corresponde então a um processo que envolve a descrição do suporte físico, no qual se encontra a provável informação, e de conteúdo, no qual se encontra materializada alguma informação e mobiliza algum conhecimento. O produto desse processo constitui um conjunto de atributos descritivos de um objeto informacional, podendo representar apenas o objeto informacional, apenas a informação contida no objeto, ou ambos (BRÄSCHER; CAFÉ, 2008). Para Bräscher e Café (2008), a representação do domínio, diferentemente do que ocorre com a informação, reflete uma visão consensual sobre a realidade representada, dentro de um determinado contexto e sob uma determinada perspectiva. Então, para representar o domínio é necessária uma análise de domínio apropriada para compreender como os diversos atores sócio-técnicos produzem e interpretam o conteúdo.

O segundo processo de interesse, de classificação da informação, não objetiva apenas encontrar e preencher os atributos que constituirão os metadados de um documento ou conteúdo, mas empreender serviços e criar produtos que suportem o trabalho de atores sócio-técnicos na realização de tarefas, tomada de decisões e julgamentos que dependam de conhecimento e informação. Esses serviços e produtos, difíceis de implementar e atualizar, devem atender às necessidades informacionais de seus usuários não só no momento da implantação, mas também no futuro, quando novos objetos informacionais, novos usuários, novas necessidades e novos conhecimentos devem também ser mobilizados (SOLOMON, 2002).

Para atender a esses requisitos, alguns esquemas de classificação aplicados a propósitos específicos incluem características e são influenciadas por fatores locais, que pertencem a um domínio ou, eventualmente, parecem pertencer exclusivamente a uma comunidade interna àquele domínio. Embora essas classificações ainda possam atender seus objetivos

originais, as características locais poluem a classificação e podem levá-la à inutilidade precocemente. Para Hjørland (1998), pesquisas em classificação, principalmente na área de Biblioteconomia e Ciência da Informação, precisam evidenciar estratégias mais gerais para a classificação, modificando-a para contextos específicos, enquanto que outras áreas do conhecimento, como a Ciência da Computação, tendem a implementar estratégias orientadas a aplicações particulares, muitas vezes baseadas em evidências empíricas e nem sempre fundamentadas teoricamente.

Hjørland (1998) pressupõe que classificações não são neutras e refletem uma visão do domínio classificado, argumentando que teorias epistemológicas como o empirismo, racionalismo, historicismo e pragmatismo devem prover uma base mais segura para a classificação dos campos do conhecimento. A primeira teoria epistemológica, o empirismo, baseia-se no conhecimento a partir da observação e tenta estabelecer uma generalização através dos dados observados. A segunda, o racionalismo, se ocupa em estabelecer a busca da razão a partir do raciocínio, da lógica e da objetividade, generalizando e reduzindo a realidade a uma única estrutura de conhecimento, onde são sintetizados todos os interesses, propósitos e perspectivas. O historicismo, por sua vez, estabelece os processos históricos e culturais como causas da transformação do universo, podendo ser descritos e compreendidos com o objetivo de conhecer a realidade humana e determinar tendências. Finalmente, o pragmatismo compreende que a natureza apresenta fenômenos que podem ser reduzidos aos seus aspectos mais úteis e práticos, sendo que os demais aspectos, caso não ofereçam qualquer vantagem para os indivíduos, são desnecessários.

Especialmente para as duas últimas correntes, o historicismo e o pragmatismo, passa a ser essencial também reconhecer o contexto da informação, dos seus atores e do seu uso. No processo de análise de domínio, o contexto corresponde à configuração que os atores sócio-técnicos apresentavam no momento da produção da mensagem contida no objeto informacional e apresentam no momento da recepção da mensagem. O contexto implica nos significados que a informação assume em diferentes cenários de acesso e uso, em diferentes momentos da história, dentro de diferentes localidades, contida em diferentes suportes físicos e estruturas de documentos, para diferentes grupos sociais e por diferentes remetentes. Abordagens para a especificação do contexto, como o domínio, o ambiente de uso ou mesmo o contexto da necessidade informacional, têm figurado mais na agenda de pesquisa em classificação como fundamental para a organização do conhecimento, tendo em vista a necessidade de se implementar sistemas que adaptem a informação às tarefas e aos problemas das pessoas (SOLOMON, 2002).

Se para a análise de domínio é fundamental garantir a especificação do contexto, a análise de conteúdo dos documentos não é suficiente para representar o domínio. De fato, “não se deve esquecer de que aos documentos devem-se agregar outras fontes, visando-se alcançar outra faceta da realidade organizacional: a que corresponde às idéias implícitas, presentes no conhecimento tácito dos membros da comunidade e que contribui para o desvelamento dessa mesma realidade” (ALVARENGA; DIAS, 2012). Neste trabalho, essas outras fontes são denominadas fontes de evidência do contexto, as quais são exploradas em diferentes áreas do conhecimento e identificadas através de diferentes abordagens.

A partir da próxima subseção são tratadas brevemente as i) evidências empíricas, ii) temporais, iii) linguísticas, iv) semânticas, v) sociais e vi) espaciais para organização e classificação da informação. Elas constituem fontes de evidência de contexto para a provável informação contida em um objeto informacional e colaboram para inserir documentos a um provável contexto, isoladamente ou em conjunto.

3.1.2 Evidências empíricas

Principalmente na área da Ciência da Computação têm sido projetadas diversas técnicas baseadas em inteligência artificial que podem contribuir com a análise de domínio. No entanto, as técnicas computacionais parecem negligenciar a natureza social, cultural e histórica da informação (HJØRLAND, 2002). Segundo Hjørland (2002), o papel da Ciência da Computação e da Biblioteconomia e Ciência da Informação são diferentes. Enquanto a Ciência da Computação tenta capturar um modelo mais generalizado dos usuários de informação, a área de Biblioteconomia e Ciência da Informação está aberta a visões alternativas e demonstra inclusive as incertezas da informação aos usuários.

Apesar dessas diferenças, em ambas as áreas muitos estudos são baseados em fenômenos puramente empíricos. Segundo Tálamo (2001), a própria atividade de classificação em Biblioteconomia e Ciência da Informação mostra-se muito empírica, com a presença de decisões arbitrárias do indexador. Um problema dos métodos empíricos é que a observação de que um certo padrão ocorre não nos permite uma generalização e o empirismo não pode garantir a continuidade daquele comportamento no tempo. O porquê da similaridade também não pode ser respondido. Uma análise de domínio menos baseada em empirismo deve levar a resultados mais duradouros, embora seja mais difícil de implementar.

Os sistemas de recuperação de informação tradicionais normalmente adotam técnicas estatísticas para indexar automaticamente a informação e para ordenar resultados de recuperação automática de informação ao usuário. Tais técnicas são baseadas em evidências empíricas tais como frequência de ocorrência de termos em documentos (frequência de termos, ou TF) e na coleção de documentos (frequência inversa de documento, ou IDF), ou co-ocorrência de termos em documentos e na coleção (SALTON; BUCKLEY, 1988). No entanto, técnicas estatísticas como TF-IDF (frequência de termo-frequência inversa de documento, do inglês *Term Frequency-Inverse Document Frequency*) apresentam limitações ao tratar de documentos com dimensões, conteúdos e contextos muito diferentes, o conjunto mais comum em empresas (HU; BANDHAKAVI; ZHAI, 2003; WU et al., 2008).

Por outro lado, métodos estatísticos continuam a ser úteis e adotados em conjunto com outros métodos mais sofisticados, dada a variedade de fontes de evidências presentes em documentos corporativos (LIU et al., 2011). O principal problema reside em identificar quais fontes de evidências empíricas são mais úteis em cada fração da coleção de documentos, reconhecendo que adotar técnicas estatísticas, genéricas para toda a coleção, pode não ser viável ao longo de todo o ciclo de vida do sistema de recuperação de informação, para todos os usuários de informação, e em qualquer contexto e necessidade de busca.

3.1.3 Evidências temporais

O tempo também representa uma variável importante na análise de domínio. Documentos e informação são veículos de um discurso que encerra seus interlocutores em um determinado intervalo de tempo, momento no qual muitas vezes o indexador e o classificador não se encontram. Esse problema está ainda mais presente quando o foco está em documentos com menor formalização, resultado do diálogo instantâneo e sem compromisso de atores sociais, mas com grande potencial de servir como mapa de conhecimento tácito e social de uma corporação (SOUTHON; TODD; SENEQUE, 2002). Tálamo (2001, p. 150) nos faz recordar que "no âmbito do simbólico, a organização da informação procede através de hipóteses e não de verdades. [...] Nesse sentido, o modelo é que goza de universalidade; os objetos empíricos devem ser avaliados na provisoriedade que lhes é própria".

De fato, mesmo supondo uma corporação em que a informação esteja estática e nenhum acréscimo ocorra, é preciso lembrar que o pressuposto de sistema fechado, em que muitos sistemas de recuperação de informação se baseiam, não é válido para usuários de informação. Mesmo que a corporação não produza mais informação ou altere seu conteúdo, o conhecimento científico continua a se desenvolver e implica em mudanças de significado e de conceitos para os usuários da informação (HJØRLAND, 2002, p. 426). Assim, mesmo um conteúdo estático apresenta mudanças de significado com o passar do tempo. Apesar disso, muitos sistemas classificatórios valorizam ou necessitam de estabilidade, resistindo às atualizações e apresentando incompatibilidade de diálogo entre antigos classificadores e atuais usuários de informação (SOLOMON, 2002, p. 25).

Para Hjørland (2002, p. 436, tradução nossa), "uma perspectiva histórica e métodos históricos normalmente proveem uma perspectiva mais aprofundada, coerente e ecológica", o que requer a avaliação da informação pela compreensão das mudanças pelas quais passam organizações, pessoas, sistemas, documentos, conhecimento e informação, ao invés de se tentar compreender a informação apenas no seu tempo de produção ou de classificação. Tecnologicamente, trata-se de algo muito distante das datas de criação e publicação de documentos, pois o foco se expande para a temporalidade do conteúdo e do significado, além da temporalidade do próprio documento.

De fato, parcela importante das fontes de evidência temporal encontra-se presente no conteúdo do documento, em formas padronizadas e não padronizadas. Formatos tais como `25 de dezembro de 2013` ou `25/12/2013` são comuns, mas compartilham espaço com formatos mais apropriados para os diferentes atores sociais (autores e leitores) que se comunicam através desses documentos. É o caso de formas ambíguas como `Natal de 2013`, uma data precisa no calendário ou um período entre novembro e dezembro para equipes de marketing e vendas; períodos bem definidos como `verão de 2013` e abreviados em relatórios como `4T2013` (para 4º trimestre de 2013); períodos imprecisos e altamente dependentes da localização geográfica como `estação chuvosa de 2013`; ou mesmo definidos através de um nome de evento com significado local (`confraternização de fim de ano` ou `lançamento do projeto X`) ou global (`atentado de 11 de setembro`, nos EUA em 2001, `crise do apagão`, no Brasil iniciada em 01/07/2001, e `acidente nuclear de Fukushima`, no Japão iniciado em 11/03/2011)

(RULA et al., 2012).

Adicionalmente, há uma fração de evidências temporais que estão associadas à contemporaneidade implícita de dois eventos, indivíduos ou mesmo documentos. O reconhecimento de dois atores sociotécnicos como contemporâneos pode depender da identificação de formas mais estruturadas de datas em documentos ou da análise de conteúdo dos documentos, onde expressões temporais tais como `antes`, `em`, `após` ou `entre` precisam ser processadas. Essa última condição depende de evidências linguísticas, as quais serão tratadas na próxima seção.

3.1.4 Evidências linguísticas

Também encontram-se na literatura trabalhos que tentam atribuir significados a termos e documentos através do reconhecimento da linguagem natural presente no documento. Por esta abordagem, significados e contexto poderiam ser reconhecidos através da leitura, intelectual ou automatizada, do próprio texto, reconhecendo elementos léxicos, ortográficos, gramaticais, semânticos e contextuais (LADEIRA, 2010).

De fato, o texto é constituído de elementos que favorecem esse reconhecimento, o qual pode ser implementado em seus vários níveis de complexidade. Tais elementos são constituídos pelo a) léxico; b) a estrutura sintagmática, composta por relações sintáticas e gramaticais; e c) a estrutura paradigmática, composta por relações lógicas prévias, ditada por convenções ou usos externos ao documento e muitas vezes denominada apenas como semântica (GARDIN, 1973, p. 147).

O reconhecimento de estruturas sintagmáticas e paradigmáticas para identificar grupos de documentos similares tem sido parcialmente alcançado por métodos empíricos baseando-se no simples reconhecimento de padrões, ao invés de explorar o significado que se constrói ou se destrói pelo jogo dos autores com o léxico. O que parece mais lógico seria o contrário, tentar reconhecer o significado só após reconhecer os gêneros textuais, as estruturas dos documentos e a terminologia usados por uma certa comunidade, condição bem semelhante a do conhecimento prévio, necessário para qualquer leitor, classificador ou indexador de informação. Essa visão parte da premissa de que “todo documento pertence a algum assunto e cada assunto tem sua própria linguagem especializada” (LADEIRA, 2010, p. 48), que é independente do idioma e portanto ultrapassa a informação provida pela sintaxe (JOYCE, 2011).

Exatamente por essa necessidade de conhecimento prévio (ou informação prévia), Hjørland (2002, p. 437, tradução nossa) aponta a necessidade de que abordagens de estudo de gêneros textuais sejam baseadas em teorias mais gerais de documentos e nos lembra que “diferentes disciplinas e comunidades de discurso desenvolvem tipos especiais de documentos como adaptações para suas necessidades específicas”. Reconhecer esses tipos, caracterizá-los e evidenciar a linguagem pela qual cada comunidade se manifesta deve suportar a construção de métodos empíricos e linguísticos mais eficazes.

Exclusivamente do ponto de vista da linguística computacional, a partir da década de 1980, as principais teorias adotadas são a teoria linguística, teoria semântica e psicolinguís-

tica, sendo as duas primeiras as mais comuns na literatura. Embora tenha havido um esforço de pesquisa continuamente voltado para a sintaxe e seu uso na atribuição de sentido às palavras (NAVIGLI, 2009), houve um aumento de investimento na teoria semântica ao longo do tempo quando se passou a integrar teoria linguística e teoria semântica, continuando como principal questão "como a sintaxe e a semântica podem ser combinadas" (LADEIRA, 2010, p. 50). Evidências de contribuições de várias áreas podem ser encontradas, como da Biblioteconomia, Ciência da Informação, Ciência da Computação e Linguística, migrando gradualmente de esforços isolados e disciplinares para aqueles mais interdisciplinares e conjuntos. Porém, historicamente, a linguística tem estado cada vez menos presente em estudos na área de recuperação de informação se comparada à área de Biblioteconomia e Ciência da Informação, à Ciência da Computação (no campo de inteligência artificial, principalmente) e à Psicologia Cognitiva (GARDIN, 1973; LADEIRA, 2010). O próprio conceito de semântica nos estudos em informação parece estar divorciado da linguística. As evidências supostamente semânticas são tratadas na próxima seção.

3.1.5 Evidências semânticas

Para além das estruturas sintagmáticas tratadas na seção anterior, estruturas paradigmáticas como fonte de evidência semântica para análise de domínio, classificação e recuperação de informação são mais exploradas na Ciência da Computação e na Ciência da Informação (LADEIRA, 2010). Isso explica-se por estruturas paradigmáticas serem compostas por relações lógicas não presentes no texto, ditada por convenções sociais ou profissionais (GARDIN, 1973, p. 147), o que exige maior conhecimento sobre usuários de informação, gêneros de documentos, fundamentos históricos e conceituais das disciplinas por trás do texto, ou, em outras palavras, sobre a organização do conhecimento.

Em coleções relativamente homogêneas, a Ciência da Computação identifica essas estruturas principalmente através de técnicas estatísticas sobre evidências empíricas. Com o desenvolvimento da *Web* social, as estruturas sociais também passaram a ser exploradas, baseando-se na presença de autores, leitores, recomendações e outras ações comuns do ambiente virtual. Também é comum a adoção de estruturas semânticas previamente organizadas, como tesauros (SALTON; BUCKLEY, 1988), ontologias (ALBANI; DIETZ; ZAHA, 2006; SOSNOVSKY; DICHEVA, 2010; FERNANDEZ et al., 2011) ou mesmo sistemas de informação estruturados. Em todos os casos, essas estruturas semânticas tentam servir como fonte de reconhecimento e desambiguação de significado (HU; SVENSSON, 2010), ou para extração (DOLBY et al., 2009) ou compatibilização de vocabulário corporativo (OMELAYENKO, 2002).

No entanto, ao arquivar, reunir ou mesclar documentos em bancos de dados (semânticos ou não), significados implícitos do contexto anterior são perdidos. É preciso que esses bancos de dados sejam elaborados de modo a enfrentar essa perda de significado ou reduzir a perda ao máximo possível (HJØRLAND, 2002). Trabalhos que tentam manter esse contexto compartilham uma visão de semântica incompleta e bastante restrita. O que muitos deles propõem gira em torno de vocabulários controlados e esquemas virtuais de compatibilização de vocabulários. Embora úteis, essas estratégias ainda se encontram muito distantes de capturar

significados para diferentes grupos de usuários e em diferentes contextos.

No contexto das tecnologias *Web*, para escala mundial ou mesmo na menor escala das *intranets* corporativas, os principais trabalhos têm apostado em esquemas que incorporam metadados e ontologias, usando XML como estrutura de dados (SOLOMON, 2002). No entanto, essas estratégias dependem da disponibilidade e riqueza dos metadados usados e de um controle de vocabulário muito maior do que tem ocorrido (LA BARRE, 2010; BUKOWSKA et al., 2012).

3.1.6 Evidências sociais

Outra faceta do discurso materializado em documentos é sua construção social, seja ela direta ou indireta. Além da presença explícita de autores do texto e outros autores citados, há presença também de leitores, os quais, em conjunto com os primeiros, constituem uma comunidade onde símbolos, significados, gêneros textuais e estruturas documentais são construídos em conformidade com a teoria construtivista social de semântica (TÁLAMO, 2001; HJØRLAND, 2002).

Algumas fontes de evidência social são usadas direta ou indiretamente em técnicas que organizam documentos em redes de autoria, coautoria, citação ou cocitação, presentes principalmente em estudos bibliométricos (VANTI, 2002) e webométricos (PAGE et al., 1999). A referência direta a indivíduos ou coletivos (equipes, departamentos, projetos em que esses participam) suporta também a recuperação de entidades sociais, ao invés de documentos, tarefa útil para alguns sistemas de recuperação de informação (GUY et al., 2012). A partir do reconhecimento das entidades sociais mais relevantes para a necessidade do usuário de informação outras entidades podem ser recuperadas, documentos inclusive.

Muito do discurso presente nos documentos é parte da explicitação do conhecimento de um domínio ou de uma organização. Porém, também importam o conhecimento tácito, as comunicações informais e as fases embrionárias dos documentos formais e finais, onde técnicas informétricas associadas a evidências sociais e de outros tipos assumem maior importância (VANTI, 2002; ALVARENGA; DIAS, 2012). Por exemplo, cientistas, funcionários e governantes não têm usado serviços de informação formais para realizar algumas tarefas e parecem preferir fontes de informação físicas ou informais, mesmo sabendo que a informação formal existe e que algum prejuízo pode ocorrer caso não a adotem (ALWIS; HIGGINS, 2001; NUNES et al., 2006; CHOO et al., 2008; O'FARRILL, 2010; MARCELLA; ILLINGWORTH, 2012). O porquê dessa predileção não é claro, mas essa administração integrada da informação formal e informal é importante para a gestão do conhecimento organizacional e da informação corporativa.

Adicionalmente, é preciso saber o porquê de parte do discurso social informal não ocorrer diretamente sobre tecnologias de informação, tornando esse tipo de comunicação parte preliminar da explicitação do conhecimento. Essa condição pode ocorrer por maior facilidade da comunicação social face a face; ou ainda pode indicar uma fraqueza das tecnologias de informação para permitir a comunicação informal, a construção colaborativa de documentos e a recuperação eficaz de rascunhos. No entanto, pode ser útil para os atores sociais investir

algum esforço no desenvolvimento metodológico e tecnológico para o tratamento do conhecimento organizacional como um todo, inclusive evidências de conhecimento tácito, potencialmente representadas por meio de informação em contexto e espalhada em vários locais da empresa. Isso pode ajudar a reduzir o desperdício de propriedade intelectual já registrada, porém classificada e indexada inadequadamente, e pode melhorar a organização da informação em geral (CHOU, 2005).

As evidências sociais podem ser especialmente beneficiadas pelas fontes de evidências temporais, tratadas anteriormente na seção 3.1.3 e espaciais, apresentadas na próxima seção, uma vez que indivíduos estão acondicionados em janelas de tempo bem definidas e, apesar de contar com certa mobilidade, também podem ser georreferenciados. Assim como as formas coletivas de se referir aos indivíduos, restrições temporais e espaciais constituem boas formas de evidenciar redes sociais que sejam contemporâneas e conterrâneas.

3.1.7 Evidências espaciais

As mensagens contidas nos objetos informacionais muitas vezes incluem uma faceta espacial, uma vez que autores, equipamentos, unidades organizacionais, negócios e usuários informacionais atuam e ocorrem em espaços geográficos. Então, é preciso georreferenciar documentos, conteúdo e consultas de usuários utilizando-se de meios que favoreçam a identificação do contexto do documento ou da coleção onde o documento se encontra. Isso ocorre principalmente pelo aproveitamento de indicadores de localidade (LEVELING; HARTRUMPF, 2008), tais como nomes de lugar, nomes de empresas, códigos de endereçamento postal ou endereços (AMITAY et al., 2004; BORGES et al., 2007), sejam esses indicadores explícitos ou implícitos (LI et al., 2006).

No entanto, muitos dos trabalhos que se utilizam de evidências espaço-temporais o fazem sem contar com o relacionamento semântico entre esses atributos e os outros tipos de evidências citados anteriormente. Com isso, resultados de avaliações apontam que o desempenho das evidências espaço-temporais é igual ou menor que aquele dos modelos clássicos de recuperação baseados em empirismo e estatística (CARDOSO; SANTOS, 2008). O porquê dessa situação é uma questão de pesquisa em aberto, mas sistemas de recuperação de informação espaço-temporais não devem travar uma disputa com os métodos clássicos, mas adicionar potencial informação espacial e temporal às tentativas clássicas de inferência.

Para Jones et al. (2008), o uso de linguagem natural e a ambiguidade de nomes de lugares, a necessidade de interpretação de relações espaciais e a necessidade de construção de forma específica de ordenação de relevância espacial são algumas das principais dificuldades em se recuperar esse tipo de informação. Para reduzir essas dificuldades, o sistema de recuperação de informação deve implementar formas de associar informação espacial a documentos e criar conjuntos de referências, como os *gazetteers*, que permitam a inferência espacial (HASSAN; JONES; DIAZ, 2009; LI; TORRES, 2009).

Trabalhos têm usado repositórios que contenham dados semiestruturados ou estruturados, como *gazetteers* ou a Wikipedia (BUSCALDI; ROSSO, 2007), ou reconhecem entidades geográficas em um conjunto restrito de documentos, como notícias de jornais, onde há

vocabulário e estrutura gramatical controlados (HASSAN; JONES; DIAZ, 2009). Um problema identificado em tais iniciativas é que não se mostram extensíveis a um conjunto de documentos onde existam diversos idiomas, gêneros de documentos ou alta imprecisão geográfica.

Adicionalmente, como os *gazetteers* são de difícil construção, também são comuns os trabalhos que tentam sua construção através da própria *Web* ou repositórios de notícias (GOUVÊA, 2009). Por outro lado, a Wikipedia começou a se mostrar como um repositório valioso para a identificação de entidades geográficas, pela existência de versões em diversos idiomas e pela atualização frequente da sua coleção. Trabalhos como de Santos et al. (2008), Overell e Rüger (2008), Overell (2009) e Alencar, Davis Júnior e Gonçalves (2010) são recentes e têm chegado a resultados eficazes ao substituir integralmente os *gazetteers* pela Wikipedia ou como formas de avaliação ou de enriquecimento dos *gazetteers* por meio do conteúdo da Wikipedia. Porém, baseiam-se muito em metodologias que adotam um repositório homogêneo e previamente classificado de documentos, como páginas *Web* de um único idioma ou notícias de jornais (GOUVÊA, 2009; HASSAN; JONES; DIAZ, 2009), algo incompatível com o ambiente corporativo. Por fim, o emprego de ontologias (CHAVES, 2009), também chamadas de geo-ontologias, tem se mostrado menos comum para o *geoparsing* e a geocodificação nos fóruns de avaliação, talvez pela complexidade no reuso e na manutenção das mesmas pelas comunidades de usuários. Outra fonte de evidência igualmente incomum para identificar o contexto geográfico baseia-se no local onde o documento encontra-se como forma de inferir o contexto espacial (AMITAY et al., 2004), quando endereços *Internet Protocol* (IP) de computadores ou mesmo endereços físicos de documentos são adotados. No entanto, essas técnicas normalmente levam a resultados imprecisos e incorretos.

3.2 Análise facetada e classificação facetada

No contexto do presente trabalho, o fundamento metodológico para a análise de domínio baseia-se na teoria da análise facetada, a qual constitui sistema de organização de conhecimento baseado em categorização de assuntos, mapeados em notação, pela qual documentos são indexados (CAMPOS, 2004). Embora termos e relações entre assuntos não sejam componentes do sistema de classificação facetado, ele provê suporte para o desenvolvimento também de outros sistemas de organização como aqueles baseados em termos, como *gazetteers* e glossários, e em esquemas de categorização, como ontologias e tesauros (CAMPOS, 2004). Como técnica, a análise facetada se mostra apropriada para um processo exploratório de análise de domínio, do qual o produto esperado é um conjunto formal de características comuns da informação corporativa.

Shiyali Ramamrita Ranganathan desenvolveu a teoria motivado pela dificuldade de representar assuntos compostos através dos sistemas de classificação enumerativos. Essa foi a mesma motivação do *Classification Research Group* (CRG), criado em 1952 no Reino Unido, ao dedicar-se ao estudo dos sistemas de classificação bibliográficos e ao adaptar a teoria de Ranganathan para torná-la menos restritiva à classificação de qualquer sistema bibliográfico (GARFIELD, 1984; BROUGHTON, 2006).

Feita a análise de um domínio, a partir de uma amostra representativa de seus docu-

mentos, tornam-se conhecidas suas características mais comuns e podem ser empreendidas tentativas de sintetizar tais características para melhor descrever e categorizar assuntos e documentos desse mesmo domínio. Pela técnica da análise de facetas, cada assunto do domínio pode ser reconhecido por variadas características e representado em diferentes perspectivas ou facetas (LIMA, 2004a, p. 58). A representação facetada dos assuntos é que permite a descrição ou indexação de documentos e sua futura recuperação a partir das várias perspectivas que os usuários têm dos assuntos daquele domínio.

Como facetas são originadas da aplicação de um princípio básico de divisão, autores como Hjørland (1998), Broughton (2006) e La Barre (2010) associam a teoria facetada ao racionalismo e a lógica como teorias epistemológicas de base. Porém, a maioria dos sistemas de classificação, principalmente aqueles baseados em facetas, também está fortemente embasada no pragmatismo, sendo orientada aos propósitos de cada classificação (HJØRLAND, 2002; CAMPOS, 2004; BROUGHTON, 2006).

Uma faceta corresponde então a uma classe de conceitos com igual relacionamento e pelo menos uma característica comum, aquela que foi usada para a divisão. Dentro de uma faceta, subconjuntos de conceitos podem ser agrupados, pela aplicação de novos princípios de divisão, constituindo subfacetas, que são novas classes de conceitos que possuem uma ou mais características comuns (LIMA, 2004a, p. 58). Essas subdivisões sucessivas, de um assunto mais geral para um assunto mais específico, formam uma estrutura hierárquica de assunto conhecida como cadeia. As classes formadas a partir de uma única característica de divisão, por sua vez, são conhecidas como renques (LIMA, 2004a, p. 62).

As facetas mais gerais do domínio são também conhecidas como categorias fundamentais, ou apenas categorias (SPITERI, 1998). Para Ranganathan, as seguintes cinco facetas poderiam servir como categorias fundamentais para qualquer domínio. São elas: *Personality* (personalidade), *Matter* (matéria), *Energy* (energia), *Space* (espaço) e *Time* (tempo), normalmente conhecidas pela sigla PMEST (GARFIELD, 1984). Essas categorias fundamentais poderiam ser então subdivididas em subfacetas ou mesmo expandidas para novas facetas em função da necessidade do classificador (GOPINATH, 1992). Este é o caso do *Classification Research Group* (CRG) que preconiza treze categorias em seus trabalhos, a saber: *Thing* (coisa), *Kind* (tipo), *Part* (parte), *Property* (propriedade), *Material* (material), *Process* (processo), *Operation* (operação), *Agent* (agente), *Patient* (paciente), *Product* (produto), *Byproduct* (subproduto), *Space* (espaço) e *Time* (tempo) (BROUGHTON, 2006). No entanto, nenhuma dessas categorias é obrigatória e as categorias fundamentais de uma classificação facetada devem ser reconhecidas pelo classificador em conformidade com o domínio a classificar, seus propósitos e sua utilidade (SPITERI, 1998).

Nenhuma das categorias citadas anteriormente é característica exclusiva da informação corporativa. Entretanto, elas são potenciais candidatas por pertencerem às entidades corporativas e não-corporativas presentes nos documentos produzidos pelas empresas. É um requisito essencial que as categorias sejam boas candidatas a atributos permanentes; ou seja, que organizem e relacionem as diversas entidades presentes na comunicação empresarial ao longo do tempo e do espaço (LA BARRE, 2010). Com isso, tais candidatas potencialmente

constituiriam um modelo de longo prazo para a organização da informação corporativa, requerendo poucas revisões e favorecendo o desenvolvimento de sistemas de recuperação de informação corporativa, suficientemente flexíveis, para integrar toda e qualquer unidade organizacional da empresa (VAKKARI; JÄRVELIN, 2005). Dessa forma, as categorias e facetas, descobertas a partir da análise facetada, seriam mapeadas em facetas de uma classificação facetada para representar a informação corporativa.

Dentro das categorias – sejam elas categorias fundamentais, facetas ou subfacetas, formando cadeias ou renques – estão os conceitos (no plano das ideias). Os conceitos são unidades de pensamento que podem ser descritas por meio de termos próprios do domínio classificado, sendo os termos representações verbais usando uma ou mais palavras da linguagem natural (no plano verbal) (LIMA, 2004a, p. 62). Embora sejam os termos os presentes nos documentos do domínio, a indexação de documentos dá-se normalmente através de uma notação que descreve seu assunto (no plano notacional), o que simplifica tanto a organização da informação quanto a futura atualização das classes e busca e recuperação da informação (SPITERI, 1998).

O assunto a ser representado pela notação é constituído por zero, uma ou mais ideias. Quando não há ideia isolada, ou isolado, como componente de um assunto, este é chamado de assunto básico. Um assunto composto é então constituído de seu respectivo componente, com um ou mais isolados, e um assunto básico de origem. O isolado, como uma ou mais ideias, só determina realmente um assunto quando dentro de um contexto, ou seja, quando combinado com um assunto. Situação análoga ocorre no plano verbal, onde um termo só possui significado dentro de um contexto ou, em outras palavras, quando também combinado com um assunto (CAMPOS, 2004). Como isolados são componentes dos vários assuntos compostos presentes na classificação facetada, cada faceta também possui seu isolado como componente. Porém, dentro das facetas também são encontrados grupos de isolados, como as subdivisões das facetas/subfacetas, os quais são chamados focos e correspondem no plano verbal aos grupos de termos associados às facetas (LIMA, 2004a, p. 61).

A teoria da análise facetada de Ranganathan (1967) constitui-se de 46 cânones, 13 postulados e 22 princípios e apresenta-se através de um texto considerado exigente para o leitor, que requer muita atenção e releituras (SPITERI, 1998). Por outro lado, modificadas pelo *Classification Research Group* (CRG) as exigências da teoria tornaram-se menos restritivas e melhor ilustradas em classificações reais. Porém, a teoria de análise facetada do CRG "não se encontra em fontes específicas, mas dispersa em vários trabalhos publicados pelos diferentes membros do grupo" (LIMA, 2004a, p. 64). Essa condição motivou o trabalho de Spiteri (1998) que pode ser sintetizado, em suas próprias palavras, como "um modelo de análise facetada consolidado e simples de seguir" (SPITERI, 1998, Sec. 2, tradução nossa).

As principais diferenças entre os trabalhos de Ranganathan e do CRG podem ser consultadas em fontes variadas como Lima (2004a), Lima (2004b), Spiteri (1998) e Broughton (2006), cabendo a esta seção apresentar apenas a visão simplificada por Spiteri (1998) e discutir os princípios assim como adotados nesta pesquisa. Na seção 3.2.1 são apresentados os procedimentos para formação de assuntos prescritos por Ranganathan; na seção 3.2.2 são

apresentados os princípios simplificados para formação de categorias e facetas; finalmente, na seção 3.2.3 são ilustradas aplicações da teoria da análise facetada em trabalhos que também objetivam melhorar o desempenho da recuperação de informação.

3.2.1 Formação de assuntos

Lima (2004a, p. 62) aponta que a noção de categoria é usada tanto na formação de assuntos quanto na formação de categorias, sendo o processo de formação de assuntos anterior ao da categorização. O processo de Ranganathan para formação de assuntos dá-se por meio de cinco métodos, i) dissecação, ii) laminação, iii) desnudação, iv) agregação e v) sobreposição, os quais são apresentados brevemente nos próximos parágrafos.

O método da dissecação (*dissecation*) consiste em dividir o universo analisado em lâminas, sendo que cada lâmina representa um assunto básico ou um isolado. Sucessivas iterações dividem as lâminas em novas lâminas de níveis diferentes. Um exemplo possível e não exaustivo do resultado da aplicação do método de dissecação é um renque com as áreas profissionais mobilizadas em uma instituição, como Contabilidade, Direito, Engenharia, Gestão de competências, Informática, Marketing e Vendas. Um segundo renque de exemplo, também não exaustivo, é resultado da dissecação da lâmina do isolado Contabilidade, que poderia ser Contabilidade de custos, Contabilidade financeira e Contabilidade tributária.

O segundo método, a laminação (*lamination*), consiste em formar assuntos compostos a partir da combinação de assuntos básicos e isolados, a partir de duas facetas. Um exemplo de assunto composto é Engenharia de embalagem, pela laminação do assunto básico Engenharia e do isolado Embalagem. Outro exemplo de assunto composto pode ser Vendas de dezembro, pela laminação do assunto básico Vendas e do isolado Dezembro.

O terceiro método é a desnudação (*desnudation*), pelo qual são identificados assuntos com grande profundidade (intenção) a partir de assuntos mais gerais ou isolados. O resultado é uma cadeia de assuntos em que a profundidade (intenção) aumenta e a extensão diminui a cada iteração do método de desnudação. Uma cadeia de exemplo é formada pela hierarquia entre os assuntos Contabilidade $\supset$ Contabilidade tributária $\supset$ Controle de registros fiscais $\supset$ Auditoria fiscal. O assunto Auditoria fiscal constitui um subconjunto de Controle de registros fiscais, que constitui um subconjunto de Contabilidade tributária, que finalmente está contido em Contabilidade.

Agregação (*loose assemblage*) é o quarto método. A agregação resulta em um assunto complexo formado por um assunto básico ou composto e por isolados. Lima (2004a) apresenta alguns exemplos de assuntos complexos dados por Ranganathan, como “Relação geral entre a Ciência Política e a Economia, Análise estatística para gerentes de ferrovias, Influência da Geografia na História”, e de isolados complexos, como “Influência do Budismo no Cristianismo” e “Diferença entre vertebrados e invertebrados” (LIMA, 2004a, p. 63). Seguindo os exemplos dos métodos anteriores, um assunto complexo possível é Contabilidade de custos da engenharia de embalagens em dezembro. Dois isolados complexos são Desempenho da auditoria fiscal em dezembro, e Propostas da engenharia de embalagens para vendas de dezembro.

Finalmente, o quinto método corresponde à sobreposição (*superimposition*). Por meio desse método são criados isolados compostos a partir de dois ou mais isolados que pertençam ao mesmo universo de isolados. Novamente o exemplo de Lima (2004a) parece útil antes de voltarmos ao contexto corporativo. Tomando o isolado professor dividido em duas características, campo de atuação e habilidade retórica:

> ele [Ranganathan] considera essas duas características como uma ideia quase isolada, de maneira que os assuntos formados pela a [sic] reunião destas duas características são ideias superpostas, como por exemplo, professor de química brilhante, professor de zoologia medíocre (LIMA, 2004a, p. 63-64).

A partir dos exemplos do método anterior, é possível produzir os isolados compostos Orçamento da engenharia de embalagens para vendas de dezembro; Orçamento do marketing para vendas de dezembro; e Orçamento da informática para vendas de dezembro.

Pela aplicação de um ou mais métodos de formação de assuntos, é identificado e analisado um número ilimitado de conceitos, para que seja identificado e analisado um número ilimitado de assuntos. Os assuntos devem servir de base para uma futura indexação, descrevendo os recursos informacionais e dando a eles um significado que é mapeado da relação entre os assuntos para o objeto sendo classificado. No entanto, a relação entre os assuntos só é obtida através da formação de categorias, facetas e subfacetas úteis para que esse mapeamento de significado ocorra – do sistema classificatório para os objetos, pelo indexador; e da necessidade de busca para os objetos classificados, pelo usuário. Os princípios teóricos para formação de categorias, facetas e subfacetas são tratados na próxima seção, 3.2.2.

3.2.2 Formação de categorias

Spiteri (1998) sintetizou os princípios do *Classification Research Group* (CRG) e os cânones, postulados e princípios de Ranganathan em três conjuntos de princípios: i) o do plano de ideias, com nove princípios, que dão suporte à síntese das ideias e características do domínio para representá-lo através de facetas; do plano verbal, com dois princípios, que estabelecem o protocolo para eleger termos significativos para representar as ideias do domínio; e do plano notacional, com quatro princípios, onde termos em linguagem natural são mapeados para uma notação pela qual ocorre a indexação. Os princípios serão apresentados e discutidos nas próximas três subseções.

3.2.2.1 Princípios do plano das ideias

O primeiro grupo de princípios, do plano das ideias, reflete na escolha de facetas que servem para a representação de assuntos do domínio e na escolha da ordem de citação das facetas e também dos isolados dentro dos renques. Os nove princípios deste plano são apresentados e explicados a seguir.

O *princípio da diferenciação* resulta na divisão de uma classe a partir de uma diferença que seja comum entre os elementos dessa classe. Foi proposto por Ranganathan como um cânone e um exemplo trivial de sua aplicação é a divisão de Seres humanos pela característica Gênero, em Feminino e Masculino (LIMA, 2004a, p. 66-67). Outro exemplo possível,

também trivial, é a divisão da classe de Atividade econômica (Economy) pela característica Setor econômico (Degree of activity), em Setor primário (Primary stage), Setor secundário (Secondary stage), Setor terciário (Tertiary stage) e Setor quaternário (Quaternary stage).

O *princípio da relevância* resulta em facetas significativas para usuários informacionais, na medida em que realmente refletem o domínio e seus assuntos e os objetivos do sistema de classificação. Foi proposto por Ranganathan como um cânone e pelo CRG como um princípio. Um exemplo da aplicação desse princípio é a divisão das classes Meninos e Meninas pela característica Grau escolar, dentro do escopo de uma classificação da disciplina Educação. A divisão poderia resultar nas subclasses Ensino infantil, Ensino fundamental e Ensino médio (LIMA, 2004a, p. 67). O exemplo da divisão de atividade econômica em setor econômico, pertencente ao escopo do domínio corporativo e dado no princípio da diferenciação, também atende ao princípio da relevância.

O *princípio da verificação* resulta em facetas a partir de características que possam ser verificadas ou mensuradas. Foi proposto como um cânone por Ranganathan e mantido como um princípio pelo CRG. Spiteri (1998) e Lima (2004a) divergem em seus pontos de vista sobre um exemplo baseado na raça de cães. Pelo exemplo, Raça é uma faceta de Cão que atende ao princípio da verificação, "uma vez que há fontes disponíveis que listam os vários tipos de raças de cães e que são reconhecidas por criadores e veterinários" (SPITERI, 1998, Sec. 5). Em termos modernos, exames de *deoxyribonucleic acid* (DNA) podem ser usados para verificar se um cão é de uma raça ou não, mesmo que criadores e veterinários possam listar raças de maneiras diferentes (LIMA, 2004a, p. 67). O exemplo da divisão de atividade econômica em setor econômico pode ser verificado por garantia literária, na literatura da área de economia, mesmo que outras formas de divisão estejam disponíveis. A divisão ilustrada é a comumente encontrada na literatura em geografia, economia e em entradas enciclopédicas.

O *princípio da permanência* resulta em facetas que refletem características permanentes da classe a ser dividida. Em outras palavras, "facetas usadas no sistema de classificação devem ser mantidas enquanto não houver mudança no propósito do sistema" (SPITERI, 1998, Sec. 3, tradução nossa). Foi proposto como um cânone por Ranganathan e foi mantido pelo CRG. Porém, Spiteri (1998, Sec. 3, tradução nossa) defende que o exemplo de Ranganathan "sugere que permanência não representa usar as mesmas facetas [facetas permanentes da classe], mas usar facetas que refletem características permanentes da entidade em questão [características permanentes da entidade]". Esta visão acabou prevalecendo em exemplos diversos encontrados na literatura e na definição do princípio da permanência do CRG. No entanto, a explicação dada (sobre o exemplo de raça de cão) por Spiteri (1998) e Lima (2004a) não parecem melhores que a explicação original de Ranganathan (sobre cor de camaleão), embora tanto o exemplo de raça de cão quanto o exemplo de cor de camaleão atendem perfeitamente o princípio da permanência. Ambos os exemplos são explicados a seguir.

No exemplo da raça de cão, "um cão Dálmata será sempre um Dálmata, assim a faceta raça representa uma característica permanente de Cão, apesar que se possa argumentar que os tipos disponíveis de raça de cão possa mudar" (SPITERI, 1998, Sec. 3, tradução nossa). Ou, "um cachorro de raça 'Dálmata' será sempre um dálmata" (LIMA, 2004a, p. 67). A explicação

parece tão incompatível com a definição dada para este princípio que Spiteri (1998, Sec. 3) tenta se desculpar: "*It is perhaps this latter quality* [características permanentes da *entidade*, ao invés de facetas permanentes da *classe*] *that is more important in this Canon, especially since it is reinforced in a similar CRG principle*".

É preciso lembrar que a teoria da análise facetada está interessada em características de ideias e assuntos e não em instâncias particulares que pertençam ao domínio e devam ser indexadas a partir dos assuntos. Assim, se as características de uma entidade ou instância podem mudar não é importante para o sistema de classificação. Se fosse, todos os exemplos anteriores e a maioria daqueles encontrados na literatura não estariam em conformidade com o princípio da permanência. Se um dálmata será sempre um dálmata não se tenta responder nesta pesquisa. Porém, um menino X no 1º grau pode estudar e ingressar no 2º grau ou, por uma mudança no sistema de ensino, ser reclassificado como no ensino médio? Um menino X, masculino, pode trocar de sexo? Um gato Y, selvagem, pode ser domesticado? A atividade Educação, no setor terciário, pode ser reclassificada para o setor quaternário? – Sim, certamente, para todas as perguntas. Nada disso é permanente para as instâncias X, Y e Educação. Por outro lado, na educação, estudantes serão sempre divididos em grau escolar? Seres humanos serão sempre divididos em gênero? Animais serão sempre divididos em habitat? Atividades econômicas serão sempre divididas em setor econômico? – Sim, acredita-se, para todas as perguntas.

O exemplo de Ranganathan de quando não usar uma característica como faceta por não atender o princípio de permanência é a cor de camaleão. "*Ranganathan argues that the facet 'colour' should not be used to divide CHAMELEONS, because these entities can change their colour to match their environment*" (SPITERI, 1998, Sec. 3). Ranganathan está correto, cor não é uma característica diferenciadora para a classe de camaleões. A melhor defesa para isso é o fato de que camaleões simplesmente não têm cor. Como seria possível reunir dez exemplares de camaleão de cor amarela? Bastaria colocar dez exemplares quaisquer em uma sala amarela, não importando sua cor no momento da coleta. Como comprovar que todos os camelões podem ser amarelos? Basta colocar todos os exemplares do planeta na mesma sala amarela. Uma vez lá, é possível torná-los verdes? Sim, basta tornar a sala verde. É possível distribuir os exemplares em cores diferentes? Sim, sempre aleatoriamente colocando-os em salas de cores diferentes. Finalmente, é possível manter em uma mesma sala de uma só cor exemplares de camaleão de cores diferentes? Não, uma vez que eles assumirão as cores uns dos outros, uniformemente, após reunidos. Assim, a Cor não é uma característica diferenciadora para a classe Camaleão por não ser possível (ou útil) dividir seus exemplares em subclasses baseadas em cor. Cor é uma característica externa ao camaleão. Para concluir, este exemplo se parece com o de dividir Seres humanos pela característica Cor da roupa. Não é possível porque Cor da roupa pertence à Roupa, não aos Seres humanos, os quais, apesar de quase sempre usarem roupas, normalmente trocam suas cores em função de ocasiões, humores e moda.

O *princípio da homogeneidade* resulta em facetas cujo conteúdo represente apenas uma característica de divisão, não permitindo sobreposição (*superimposition*) de características que deveriam estar em diferentes facetas. O estabelecido por este princípio é atendido

pelo cânone da concomitância de Ranganathan, mas foi proposto isoladamente como princípio pelo CRG. Um exemplo de sua aplicação pode ser derivado do exemplo do domínio corporativo. Atividade econômica pode ser dividido por setor econômico, em Setor primário, Setor secundário, Setor terciário e Setor quaternário. Setor quaternário pode ser subdividido em Pesquisa, Desenvolvimento e Educação. Pesquisa poderia ser subdividida em Pesquisa básica e Pesquisa Aplicada. Porém, por este princípio, Educação não poderia ser subdividida em Educação à distância, Educação básica, Educação especial, Educação indígena, Educação infantil, Educação presencial, Educação profissional, Educação superior e Pós-graduação, por exemplo. Isso porque nesse renque interno à Educação estaríam presentes características – modalidades de ensino, tecnologias de ensino, metodologias de ensino e níveis educacionais – que deveriam estar acomodadas em diferentes subdivisões.

O *princípio da exclusividade mútua* é complementar ao princípio da homogeneidade e resulta em subclasses formadas por uma e apenas uma característica da classe de origem, sem que a mesma característica esteja presente em mais que uma subclasse. O estabelecido por este princípio é atendido pelos cânones da concomitância e da exclusividade de Ranganathan, mas foi proposto isoladamente como princípio pelo CRG. Do exemplo do princípio anterior, podemos dizer que Educação básica dividida pela característica nível educacional em Educação infantil, Ensino fundamental, Ensino médio e Ensino técnico não é permitido. Ensino médio e Ensino técnico são equivalentes quanto ao nível educacional. Para atender ao princípio seria necessário dividir Educação básica em Educação infantil, Ensino fundamental e Ensino médio. Finalmente seria possível subdividir Ensino médio em Ensino técnico integrado, Ensino técnico com concomitância externa e Ensino técnico subsequente.

O *princípio das categorias fundamentais* define que não há categorias obrigatórias para qualquer domínio ou assunto e que as categorias fundamentais devem ser determinadas em função do domínio e dos objetivos do sistema de classificação (SPITERI, 1998). Foi proposto apenas pelo CRG, diferindo do postulado das cinco categorias fundamentais de Ranganathan. Segundo Spiteri (1998), a maioria dos sistemas de classificação facetada e tesauros consultados usam a abordagem do CRG na escolha das categorias fundamentais, um conjunto de treze categorias apenas sugeridas, mas que acabam muitas vezes sendo assumidas arbitrariamente ou mecanicamente. As categorias fundamentais de Ranganathan (PMEST) e do CRG já foram tratadas no início desta seção, 3.2.

Os dois últimos princípios deste plano “coordenam a organização dos focos dentro de suas respectivas facetas e, consequentemente, a ordenação destes [sic] focos no renque” (LIMA, 2004a, p. 68). O *princípio de sucessão relevante* resulta em uma ordem de citação das facetas que seja relevante para o objetivo do sistema de classificação e para seus usuários. Spiteri (1998) apresenta alguns modelos de ordenação compatíveis com as propostas de Ranganathan e do CRG, não apresentados nesta seção. Por todo o capítulo, as categorias fundamentais de Ranganathan e do CRG são listadas na ordem original dos autores; e outras facetas são ordenadas alfabeticamente. A exceção é o renque Educação infantil, Ensino fundamental e Ensino médio, ordenado cronologicamente. O *princípio de sucessão consistente* resulta em facetas consistentemente ordenadas em todo o sistema de classificação, sofrendo mudança na ordenação se e somente se houver mudança no propósito do sistema de clas-

sificação. Como ilustração desse último princípio, não há mudança do método de ordenação dos renques de facetas neste e nos próximos capítulos da tese, simplesmente por não haver motivo para fazê-la, ficando como padrão a alfabética.

3.2.2.2 Princípios do plano verbal

Os próximos três princípios são específicos para o plano verbal, isto é, versam principalmente sobre o reconhecimento e atribuição de significado à terminologia usada no sistema de classificação.

O *princípio do contexto* resulta na atribuição de significado para um termo em função da sua posição na estrutura do sistema de classificação, emprestando parte do contexto das classes nas quais se encontra inserido e emprestando parte do seu próprio contexto para as subclasses a partir da posição onde é isolado. Foi proposto por Ranganathan como cânone do contexto, sem correspondência na teoria do CRG, mas ainda assim mantido no modelo simplificado de Spiteri (1998). A utilidade do princípio está principalmente na resolução de significado de homógrafos, como economia que pode ser a ciência social (se estiver dentro de Disciplina), o conjunto de atividades desenvolvidas pelas instituições visando a produção, distribuição e o consumo de bens e serviços (se for dividida em setores econômicos), ou o resultado de uma operação eficiente sobre recursos (se estiver dentro da estrutura de resultados de processos).

O *princípio da terminologia usual* resulta no reconhecimento das formas mais comuns de se referir às entidades, às características e aos assuntos do domínio. Isso requer frequentes revisões do sistema de classificação procurando atualizá-lo ao desenvolvimento tecnológico, científico e terminológico que ocorre continuamente dentro daquele mesmo domínio. Uma das principais vantagens da classificação facetada é a maior flexibilidade para atualização gradual e contínua. Baseado no cânone da terminologia corrente (*currency*) de Ranganathan, não possui princípio correspondente no modelo do CRG. Um exemplo do não atendimento do princípio seria a adoção de um termo historicamente muito antigo para Gestão de pessoas, que evoluiu, em uma história mais recente, de Gestão de mão de obra, passando por Gestão de recursos humanos, até Gestão de pessoas. Ou ainda Adestramento, sinônimo de Treinamento e de Capacitação, adotado até a década de 1940. Caso a terminologia não seja a mais comum, certamente haverá uma incompatibilidade entre as visões do indexador e do usuário informacional.

O cânone da reticência de Ranganathan não tornou-se um princípio no modelo do CRG ou no modelo simplificado de Spiteri (1998), porém merece ser explicado como parte do princípio da terminologia usual, embora esteja em conflito com este princípio, como definido por Spiteri (1998). Pelo cânone da reticência a terminologia usada não pode refletir um viés da visão de um grupo de classificadores ou qualquer preconceito. Spiteri (1998) defende que o conflito desse cânone com o princípio da terminologia usual reside na possibilidade de que a terminologia usual seja preconceituosa, sendo que adotar a politicamente correta significaria reduzir a expressividade do sistema de classificação. Também apresenta um exemplo de preconceito de gênero existente no Canadá pelo termo *Fishermen* e a tentativa de torná-lo

politicamente correto pela substituição por *Fishing people*, mesmo sendo o primeiro termo o mais frequente.

Apesar de ser simples resolver o problema de *Fishing people* na própria estrutura do sistema de classificação, há problemas mais difíceis de resolver em um sistema de classificação e que devem ser tratados com muito cuidado. É o caso de características tais como Cor da pele (ainda presente em muitos sistemas de informação de trabalho no Brasil, inclusive em livros de registro de empregados) e Raça (algo em desuso para seres humanos, apesar de ser uma característica provavelmente permanente para animais). No Brasil, estas características são tão sensíveis que normalmente são preenchidas livremente por autodeclaração, não podendo ser verificadas formalmente. Como não atendem o princípio da verificação, normalmente não são boas candidatas a facetas. Há outras características que margeiam a ilegalidade como Bom pagador, Mau pagador, Boa índole, Má índole, Honesto e Desonesto, todas para a classe de Ser humano. No contexto de um sistema de recuperação de informação corporativa é muito importante que o cânone da reticência seja atendido, seja para evitar constrangimentos para o próprio usuário da informação quanto para evitar tomada de decisão sobre uma informação com viés.

3.2.2.3 Princípios do plano notacional

Finalmente, os quatro próximos princípios determinam o funcionamento da notação do sistema de classificação, resultado da compilação dos termos do plano verbal, e principal linguagem de trabalho do indexador e do usuário da informação.

O *princípio de sinônimo*, proposto por Ranganathan, assegura que um dado assunto é representado apenas por um único número de classe da notação (SPITERI, 1998). Complementar ao anterior, o *princípio de homônimo*, também proposto por Ranganathan, assegura que um número de classe da notação representa apenas um único assunto (SPITERI, 1998).

Proposto apenas pelo CRG, o *princípio da ordem de fichamento* determina que a notação deve refletir a ordem de fichamento dos assuntos no sistema de classificação. Segundo Spiteri (1998), este princípio é particularmente útil para permitir que o usuário da informação possa seguir a ordem do sistema de classificação ao buscar por uma informação através da navegação, seja em prateleiras ou em catálogos classificados. Finalmente, também proposto apenas pelo CRG, o *princípio da hospitalidade* estabelece que o sistema de classificação e principalmente a notação devem favorecer a inclusão de assuntos, facetas e isolados em qualquer ponto do sistema de classificação, sendo que o custo da mudança deve ser o mínimo para a notação (SPITERI, 1998).

Apresentados os principais conceitos e princípios da teoria da análise facetada, a próxima seção, 3.2.3, ilustra alguns trabalhos que têm usado facetas diretamente ou indiretamente para melhorar o desempenho da busca, da recuperação, da ordenação e da visualização de resultados em sistemas de recuperação de informação.

3.2.3 Uso de facetas na busca e na recuperação de informação

A adoção da teoria da análise facetada é ampla na área da Biblioteconomia e Ciência da Informação principalmente para o projeto de sistemas de classificação e tesauros (BROUGHTON, 2006). Adicionalmente, em outras áreas do conhecimento tem havido uma crescente adoção das técnicas de análise facetada ou da abordagem de facetas, mesmo que os trabalhos não citem formalmente Ranganathan e o *Classification Research Group*.

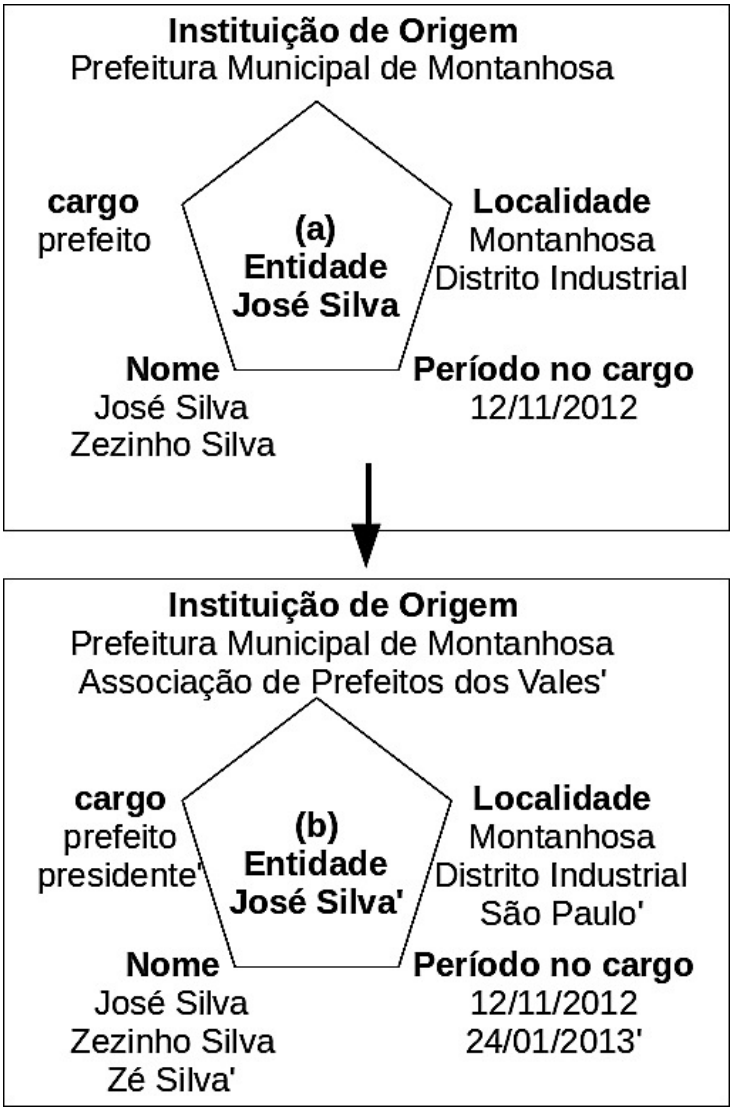
Alguns trabalhos são ilustrados nesta seção como uma motivação tecnológica para a adoção da teoria em implementações de sistemas de indexação automática e sistemas de recuperação de informação, embora a maioria dos trabalhos externos à área da Biblioteconomia e Ciência da Informação careçam de uma visão mais aprofundada do que seja a análise facetada, uma faceta e dos benefícios que a teoria pode prover para seus problemas de tratamento da informação (LA BARRE, 2010). A figura 2 serve para associar esses conceitos, assim como usados nesta pesquisa, a exemplos próprios do domínio corporativo.

Um exemplo para o relacionamento entre entidades, facetas e termos, pode ser ilustrado por meio de uma ata de reunião, um típico documento de registro em instituições. O evento reunião teria acontecido em uma data, em um município e com participantes reconhecidos por nomes completos e suas respectivas instituições de origem. Um indivíduo denominado “José Silva” é uma entidade presente no documento, ilustrada na figura 2 como um pentágono. Sua instituição de origem é a “Prefeitura Municipal de Montanhosa”, onde está no cargo de “prefeito”. `Nome`, `Instituição de origem`, `Cargo` e `Localidade` são facetas importantes para as entidades relacionadas a pessoas; e os termos “José Silva”, “Prefeitura Municipal de Montanhosa” e “prefeito” estão associados às facetas na própria construção do texto da ata. O termo “Montanhosa” estaria associado à `Localidade` por dedução a partir da sua instituição de origem ou da localidade onde ocorreu a reunião. Facetas (em negrito) e seus respectivos termos constituem diferentes dimensões do pentágono (a) da figura 2.

Reconhecer esse conjunto de facetas comuns a maioria dos documentos é essencial para a organização automática da informação corporativa e para a interoperabilidade entre diferentes repositórios. Partindo do mesmo exemplo anterior, ao ser incluída uma nova faceta temporal `Período no cargo`, mesmo a data de início não tendo sido descrita na ata, sabe-se que na data da reunião “José Silva” era o prefeito em exercício. Assim, a `data de ocorrência` associada à reunião pode ser tomada como início do cargo de todas as entidades citadas no documento, inclusive da entidade “José Silva”.

Essa associação entre localidades, datas, processos e entidades deve ser dinâmica e depende do reconhecimento de características espaciais, temporais, descritivas e lógicas. Nesse exemplo, a associação entre `data de ocorrência` da reunião e `período no cargo` de “José Silva” não é definitiva e vigora até que um novo documento evidencie uma data diferente para esta faceta, como é demonstrado na figura 2, em (b). Um documento de uma nova reunião ocorrida em São Paulo teria produzido as mudanças refletidas na entidade da figura 2 (b), onde uma nova possibilidade de nome ocorre, a presidência de uma associação é registrada e o cargo de prefeito continua vigorando. A importância que esses novos termos assumem para cada faceta depende de muitas variáveis, como tempo, espaço, atores sociais envolvidos, a

Figura 2 – Exemplo de associação entre entidades, facetas e termos. Entidades permanentes apresentam facetas permanentes que podem estar associadas a termos diferentes ao longo do tempo. É o caso do exemplo exposto pela transição entre as entidades (a) e (b). Os termos diferentes são marcados por um apóstrofo.



Fonte: elaborada pelo autor

frequência e o uso dos novos e antigos termos nas mensagens que os documentos portam.

A organização facetada da informação também torna mais flexíveis o *ranking* e a visualização de resultados de busca. Na perspectiva de um usuário de informação, uma busca textual por "negócios em montanhosa" poderia muito bem retornar documentos de projetos em que "José Silva" seja citado mesmo que os termos "Montanhosa" ou "Prefeitura Municipal de Montanhosa" não estejam presentes. Adicionalmente, a própria desambiguação do termo montanhosa, uma cidade onde a organização possui unidade ou um adjetivo que pode pertencer a muitas entidades na coleção, é uma atividade necessária dentro do sistema de recuperação. Esse tipo de inferência espacial seria naturalmente realizado por um classificador humano experiente. Para uma busca "negócios com prefeitura de montanhosa", também seria natural que projetos recentes e antigos fossem recuperados, desde que incluam os nomes de "José Silva" ou de antigos prefeitos de Montanhosa, mesmo que os termos "Prefeitura Municipal de Montanhosa", "prefeito" ou "Montanhosa" não estejam explicitamente presentes no conteúdo do documento. Um buscador experiente poderia expandir a expressão de busca e incluir o nome dos vários prefeitos. Porém, uma organização de informação que inclua facetas espaciais, temporais, descritivas e lógicas também poderia subsidiar esses tipos de inferência, mas com a vantagem de considerar a temporalidade da informação, dos documentos e de cada prefeito-entidade indexado. Ou seja, não bastaria o modelo ser facetado para suportar buscas e recuperação de informação eficientes; o modelo precisa de facetas e tratamentos que realmente ajudem a evidenciar o contexto de entidades e documentos ao longo de todo seu ciclo de vida.

Uma motivação tecnológica refere-se à visão facetada do usuário. Uma visão do esforço de busca do usuário informacional é a de que, ao invés de fornecer termos de interesse, o usuário normalmente fornece facetas de interesse. Por essa visão, caso o usuário de informação forneça um termo para espaço geográfico, como *São Paulo*, seu interesse ultrapassa a existência do termo ou de uma notação para São Paulo, acrescentando à busca uma possibilidade de explorar a geografia do seu assunto de interesse. Na existência de um número como 1930 dado pelo usuário, se tem possivelmente uma faceta temporal pela qual a busca pode ser contextualizada e para qual até mesmo a terminologia do sistema de classificação precisa se adequar (HONG, 2006; BORGES et al., 2007; BAEZA-YATES; RIBEIRO-NETO, 2011). Como exemplificado na seção anterior, se 1930 é um ano e São Paulo uma cidade ou estado, um documento com conteúdo que inclua *adestramento* e *mão de obra* refere-se provavelmente à capacitação dos funcionários das indústrias, o que requer um mapeamento de significado entre dois sistemas de classificação muito diferentes.

Se explicar a intenção da busca em poucas palavras já parece difícil, mais difícil é a mesma intenção para espaços geográficos e janelas de tempo muito diferentes do espaço e do tempo da maioria dos documentos da coleção; e ainda mais difícil quando um componente de *software* é o responsável por indexar e "interpretar" a necessidade de busca a partir de ideias e conceitos, recuperar documentos que possuem pouco mais que palavras e apresentar os resultados para o usuário de informação (HONG, 2006).

A dificuldade é ainda maior no contexto de um sistema de recuperação de informação

para documentos da *World Wide Web* (WWW), ou um sistema de recuperação de informação de Internet. Na Internet estão documentos que versam sobre os assuntos mais diversos, em linguagem natural, sem metadados, nos mais diversos idiomas e linguagens especializadas (HONG, 2006). Neste caso, o sistema de classificação deve atender aos seguintes requisitos:

> encontrar similaridade sem apenas encontrar padrões sem contexto; ser preciso e altamente descritivo; ser fácil de adicionar, apagar e atualizar classes e vocabulário sem a necessidade de reclassificar; deve suportar documentos digitais com uma informação que se expande continuamente (HONG, 2006, p. 46, tradução nossa).

Para grupos específicos de documentos na Internet, ou para comunidades específicas de usuários, existe perspectiva de que as tecnologias de Web semântica e de dados ligados (*Linked data*) farão maravilhas. Boa parte de seus trabalhos são baseados em metadados construídos voluntariamente por usuários, sem um compromisso adequado com o controle terminológico para a manutenção de longo prazo dos repositórios.

Facetas, ou os atributos que são chamados de facetas, normalmente servem principalmente ao propósito de navegar parte desses grandes repositórios (OREN; DELBRU; DECKER, 2006). O reconhecimento de facetas facilita também a navegação nos resultados uma vez que há um refinamento possível para cada faceta, algo não facilmente alcançável em um sistema de recuperação de informação baseado em palavras do texto completo (SACCO, 2007; AUER; LEHMANN; HELLMANN, 2009; GIRGENSOHN et al., 2010; MACULAN; LIMA, 2011; PONTES; LIMA, 2012). Porém, "sem um compromisso com vocabulários controlados, resta pouca esperança de que as tecnologias da Web semântica alcancem seu fim" (LA BARRE, 2010, p. 269, tradução nossa), principalmente fora do escopo das coleções de documentos altamente estruturados como é o caso de bibliotecas de teses e dissertações, de patentes e de bancos de dados.

Hong (2006, p. 47) defende o uso de classificação facetada em sistemas de recuperação de informação muito heterogêneos:

> *faceted classification focuses on the essential and constant characteristics/facets, which is useful for micro-grained rapidly changing information repository. It can be used to create deeper and more complex knowledge structures by exploring variants of combination (HONG, 2006, p. 47).*

A simplicidade dessa estrutura favorece os sistemas de recuperação de informação em que usuários devem buscar informação sobre entidades, espaços e tempos conhecidos (TODA et al., 2008; LIU et al., 2012), especialmente no momento da consulta (ANDRADE; SILVA, 2006). Este é o caso principalmente do contexto dos sistemas de recuperação de informação geográficos, em que espaço e tempo, duas facetas importantes em quase todo sistema de informação, são bem melhor estruturados e onde estão presentes alguns métodos de ordenação espaço-temporal.

Como exemplo, Ehlen, Zajac e Rao (2009) e Toda et al. (2008) descrevem aplicações para a ordenação baseada em múltiplas facetas, como a localização de documentos que pertençam à região de interesse do usuário, sendo que a relevância da faceta espacial é dada

pela menor distância do ponto onde o usuário se encontra, o que é particularmente útil em serviços de localização e serviços móveis. A faceta temporal também é considerada em muitas aplicações em que novidades sobre um certo fenômeno, ao invés de relatos históricos, são preferíveis. É o caso de repositórios de notícias, serviços de monitoramento de catástrofe ou de trânsito. De fato, o interesse é por um método que consiga mensurar a importância do contexto geográfico, temporal ou temático em tempo de consulta (ANDRADE; SILVA, 2006) e responda com eficiência ao usuário de informação (YU; CAI, 2007; GARCÍA-CUMBRERAS et al., 2009).

Uma motivação metodológica refere-se à flexibilidade da classificação facetada. Como a classificação facetada é flexível e, portanto, adequada para modelar domínios em contínua expansão e mudança, mesmo os sistemas de classificação inicialmente implementados sem um sólido embasamento teórico podem ser beneficiados (HONG, 2006; LA BARRE, 2010). Ou, em outras palavras, independentemente da qualidade da abordagem implementada pelos sistemas de recuperação de informação que se encontram na literatura, a teoria da análise facetada posta em prática no contexto desses sistemas e esses sistemas postos em avaliação sob a lente da teoria da análise facetada constitui um passo importante para a área.

Na próxima seção são investigadas as principais pesquisas sobre avaliação de modelos de domínio e sobre avaliação de sistemas de recuperação de informação, do desempenho dos processos de indexação e da recuperação de informação.

3.3 Avaliação de sistemas de recuperação de informação

Como a análise de domínio deve produzir um produto, um modelo de alto nível do domínio em estudo, após validado ele precisa ter sua utilidade avaliada. Também há muita diversidade de métricas de avaliação de utilidade desse tipo de instrumento, sendo que diferentes métricas refletem os objetivos do modelo, exatamente como acontece com a escolha dos métodos que são executados no processo de análise de domínio. Como o modelo resultante desta tese objetiva favorecer a atividade de recuperação de informação, sua avaliação se dá indiretamente, pela avaliação do desempenho da própria recuperação de informação.

A presença de uma coleção de referência favorece a comparação de sistemas de recuperação de informação e a definição de valores que sirvam como base de comparação. O contrário, quando experimentos são realizados em coleções privadas, a publicação de resultados se baseia muitas vezes em informação protegida, não disponível publicamente e muitas vezes sensível. Há muitas metodologias de avaliação de sistemas de recuperação de informação (VOORHEES, 2002; BUCKLEY; VOORHEES, 2004; BUCHER et al., 2005; SAKAI; KANDO, 2008; PETERS et al., 2008), algumas delas bem aceitas na indústria e na academia. Porém, há limitações para avaliação de sistemas corporativos pelas mesmas metodologias, dadas suas especificidades (CRASWELL; VRIES; SOBOROFF, 2005).

Embora a trilha *Enterprise* da *Text Retrieval Conference* ofereça uma coleção de referência para sanar parte das limitações de avaliação; ao ser baseada apenas em documentos da *Web* pública de uma empresa ela também se mostra insuficiente para avaliar a busca por

recursos de informação reais (BAILEY et al., 2007b). Porém, a dificuldade de melhorar o *corpus* de avaliação está na dificuldade de expor dados sensíveis de uma empresa, seus clientes, funcionários, fornecedores e suas estratégias de negócio. Mesmo uma empresa pública ou governo conta com informação sensível, protegida por lei, não sendo necessariamente um bom candidato a fornecedor de dados de avaliação. Por essa razão, muito do esforço de atingir os objetivos da trilha *Enterprise*, que aconteceu até o ano de 2008, foi transferida para uma nova trilha, *Entity Search* (ou, Busca por Entidades), que aconteceu até o ano de 2011, vislumbrando reconhecer entidades humanas e não-humanas e em diferentes escalas, das intranets de empresa à toda a Web (BALOG et al., 2008).

As técnicas e sistemas implementados têm sido confrontados com outros sistemas sob o mesmo prisma de avaliação, onde são encontradas várias metodologias. No entanto, os diferentes prismas e metodologias normalmente valorizam mais algumas fontes de evidências do que outros. É o caso da trilha *GeoCLEF*, do *Cross-Language Evaluation Forum* (CLEF), constituindo um *framework* através do qual sistemas que utilizem linguagens europeias podem ser comparados sob os mesmos critérios e com a mesma massa de dados (DOMÈNECH, 2007; CARDOSO; SOUSA; SILVA, 2008; ANASTÁCIO, 2009), visando reconhecer evidências espaciais e linguísticas. Metodologia similar, porém específica para sistemas em língua portuguesa, é denominada Avaliação de Reconhecimento de Entidades Mencionadas (HAREM) (SANTOS; CARDOSO, 2007).

Outra estratégia de avaliação baseia-se na adoção direta dos usuários do sistema de recuperação de informação. É o caso do *Open Directory Project* (ODP), onde são realizadas as mesmas atividades de avaliação e os resultados são comparados com aqueles anotados manualmente por voluntários (AMITAY et al., 2004), e as metodologias de avaliação que empregam os próprios usuários do sistema como forma de avaliação mais criteriosa dos componentes de consulta e da qualidade percebida da resposta (BUCHER et al., 2005; YU; CAI, 2007; PONTES, 2013). Em todas as metodologias de avaliação são adotadas métricas muito próximas daquelas adotadas para sistemas convencionais, sendo que alguns trabalhos sugerem a necessidade de estabelecer metodologias mais adequadas para alguns tipos específicos de fontes de evidência e de coleções (DOMÈNECH, 2007; SAKAI; KANDO, 2008).

Um fórum que imponha o reconhecimento automático de diversos tipos de evidência, como espacial, temporal, social, linguístico e temático, por exemplo, não está disponível atualmente. Sua ausência, apesar de dificultar a tarefa de avaliação de sistemas de recuperação de informação, não representa o maior desafio de avaliação. O maior desafio continua a ser definir critérios formais de relevância, coleções de referência e métricas empíricas de desempenho adequados para diferentes atores sociais e contextos de uso. Por essa razão, esta pesquisa adota os resultados de Bailey et al. (2007b) e Balog et al. (2008) como base de comparação, mas não pode dispensar uma avaliação complementar, com coleção privada e usuários reais, para garantir que os resultados realmente representem soluções de recuperação para situações mais próximas da realidade corporativa.

Os procedimentos metodológicos para avaliação desta pesquisa sobre a coleção privada deve diferir pouco daqueles adotados por Bailey et al. (2007a). A principal diferença é

a perspectiva interpretativa adotada neste estudo; a importância dada à análise de domínio como componente formal da construção da coleção de referência e do modelo do domínio; o estabelecimento dos contextos de uso mais úteis para os usuários de informação; e o reconhecimento dos gêneros textuais e das linguagens adotadas por usuários e produtores de informação corporativa (LYKKE-NIELSEN, 2011).

4 Análise de domínio corporativo

“Não se possui o que não se compreende”.
Johann Goethe

Este capítulo trata da caracterização do domínio corporativo e objetiva explicitar quais características da informação corporativa são comuns a diferentes instituições. Dessa forma, o capítulo apresenta uma proposta de generalização do domínio corporativo e ilustra algumas características descobertas a partir da análise de domínio utilizando as técnicas de análise de assunto e análise facetada, empreendidas em cada uma das coleções de documentos pesquisadas. Dentro do domínio corporativo, somente através de uma análise de domínio rigorosa pode-se tentar responder se as facetas adotadas por empresas diferentes apresentam alguma semelhança.

Os procedimentos metodológicos da análise de domínio foram executados como segue: i-a) análise de assuntos e o levantamento de termos mais significativos em cada uma das coleções de documentos; i-b) formação de assuntos pela leitura e pela aplicação das técnicas de dissecação e desnudação; ii) categorização dos assuntos pela aplicação da técnica da análise facetada em cada coleção de documentos; e iii) consolidação das categorias, facetas e subfacetas comuns às duas coleções de documentos.

Os procedimentos de análise, formação e categorização de assuntos (i-a, i-b e ii, respectivamente) foram executados em cada coleção, isoladamente, pelo autor deste trabalho. Os dois primeiros são reportados na seção 4.1 e a categorização é reportada na seção 4.2. Todos os assuntos devidamente categorizados estão disponíveis nos anexos **??**, **??** e **??**.

Por último, a análise preliminar do domínio corporativo é encerrada pelo procedimento de consolidação das facetas comuns às coleções e de avaliação do grau de similaridade que tais facetas supostamente apresentam no domínio. Seus resultados são apresentados e discutidos na seção 4.3.

4.1 Análise de assuntos e formação de assuntos

A análise de assuntos constitui a análise intelectual dos documentos de cada coleção com o propósito de interpretá-los e de representar seus assuntos por meio de termos em linguagem natural. Em seguida, sobre os assuntos resultantes foram aplicadas as técnicas de dissecação e desnudação para formação de assuntos adicionais. O processamento da análise e formação de assuntos foi dividido em três fases e ocorreu linearmente sobre as coleções particular e pública, nessa ordem.

A análise de assuntos foi realizada sobre duas coleções de documentos provenientes de duas empresas, o Centro Federal de Educação Tecnológica de Minas Gerais (CEFETMG) e a *Commonwealth Scientific and Industrial Research Organisation* (CSIRO). Como as cole-

ções pertencem a empresas diferentes, sua comparação dá-se pelos assuntos corporativos que ambas tratam. Isto é, dois assuntos, mesmo representados por termos distintos, podem ser considerados idênticos por uma avaliação intelectual do classificador. É o caso do assunto país, representado igualmente por Brasil ou *Australia*; de instituição, representado por CEFETMG ou CSIRO; e de serviço, representado por Curso de graduação ou por *Research*. Por outro lado, um mesmo termo pode referir-se a assuntos diferentes para cada empresa. Um exemplo é o termo *site* que denota um *Web site* ou um local de funcionamento da empresa para a CSIRO, enquanto só representa o primeiro assunto para a empresa CEFETMG.

4.1.1 Primeira fase de processamento: documentos da coleção particular

A primeira fase de processamento da análise de assuntos ocorreu sobre a coleção particular de documentos que pertencem à empresa CEFETMG. A coleção particular foi produzida por um gerente da empresa como um conjunto significativo e útil de documentos para o trabalho rotineiro e para tomada de decisão. Esses documentos estavam organizados em repositórios, sendo possível encontrar réplicas de um documento em um ou mais repositórios. Os repositórios mobilizados constituem a principal fonte de informação que gerentes e usuários de informação têm usado na empresa entre os anos de 2012 e 2013; portanto essa coleção é mais eficaz que uma amostra aleatória do arquivo corrente para o propósito desta pesquisa. A tabela 1 apresenta os repositórios a que pertencem os documentos com a quantidade de documentos e páginas oriundos de cada um.

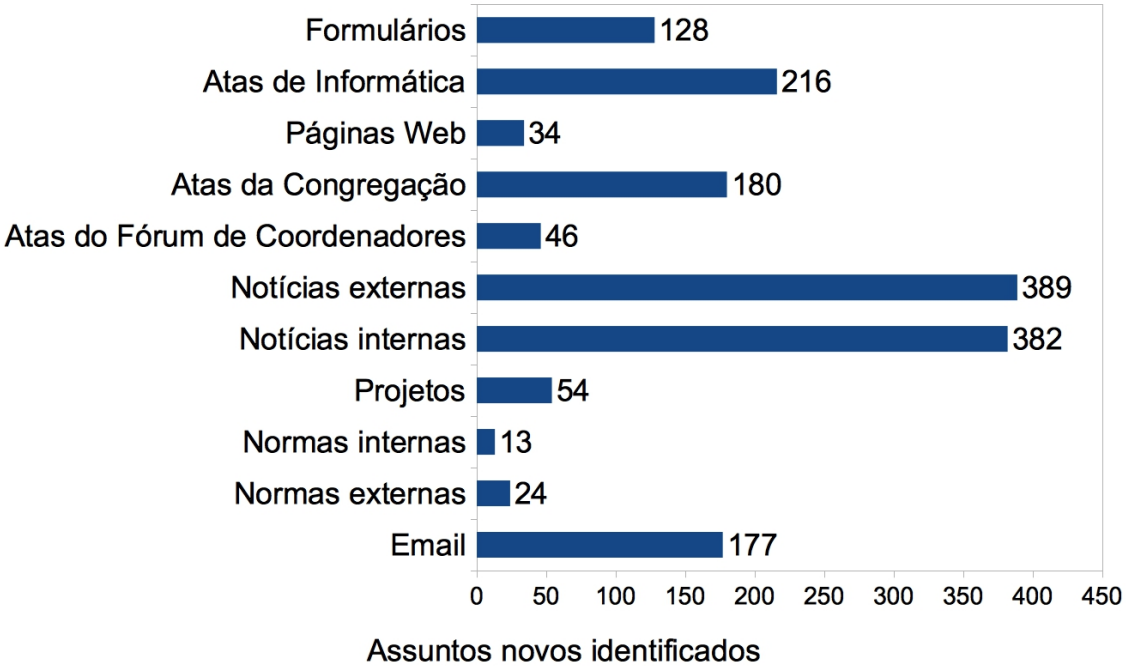
Tabela 1 – Composição da coleção particular

	Documentos (D)	**Páginas (P)**	**P/D**
Formulários	9	11	1,2
Atas de Informática	11	28	2,5
Páginas Web	5	5	1
Atas da Congregação	55	124	2,3
Atas do Fórum de Coordenadores	14	45	3,2
Notícias externas	210	210	1
Notícias internas	384	384	1
Projetos	4	426	106,5
Normas internas	6	107	17,8
Normas externas	1	109	109
Correspondências	606	606	1
Total	**1305**	**2055**	**1,6**

Fonte: Elaborada pelo autor.

Um único profissional da área, autor desta tese, leu 1305 documentos produzidos entre os anos de 2007 e 2013 e anotou os assuntos presentes na medida em que foram identificados. Nessa etapa, os assuntos foram anotados através do termo em linguagem natural que melhor o representava, empregando a terminologia adotada pela instituição em pelo menos um documento da coleção. O documento onde o assunto/termo aparecia pela primeira vez também foi anotado, mas esse dado não é disponibilizado nesta tese por violar o acordo de si-

Figura 3 – Novos assuntos descobertos em repositórios da coleção particular



Fonte: elaborada pelo autor

gilo. O objetivo é evitar constrangimentos por provável informação, e.g.: o termo Joseph ocorre pela primeira vez em um processo disciplinar.

Após a leitura de cada documento e a anotação dos novos assuntos descobertos, foram aplicados os métodos de dissecação e desnudação para formação de assuntos adicionais.

Através do método de dissecação se divide o universo analisado em lâminas, sendo que cada lâmina representa um assunto básico ou um isolado. Por meio de um processo iterativo, as lâminas são dividas em novas lâminas de níveis diferentes. Um exemplo de formação de assuntos pela aplicação do método de dissecação é o renque Manhã, Noite e Tarde para turnos de trabalho.

Através do método de desnudação são identificados assuntos com grande profundidade a partir de assuntos mais gerais ou isolados. O resultado é uma cadeia de assuntos em que a profundidade aumenta e a extensão diminui a cada iteração do método. Um exemplo de formação de assuntos pelo método de desnudação, é a cadeia Ensino $\supset$ Ensino médio e profissional técnico de nível médio $\supset$ Ensino técnico. Ou seja, Ensino técnico é subconjunto de Ensino médio e profissional técnico de nível médio, que está contido em Ensino.

Os demais três métodos de formação de assuntos, laminação, agregação e sobreposição, não foram utilizados na amostra, uma vez que foi usada a linguagem natural presente na amostra, inexistindo vocabulário controlado para controle terminológico.

A figura 3 mostra a ordem de processamento, do repositório de formulários ao reposi-

tório de Emails, e o número de novos assuntos descobertos em cada repositório da coleção particular. Porém, caso outra ordem de processamento fosse seguida o total de assuntos descobertos continuaria o mesmo. A análise e formação de assuntos na coleção particular consumiu o tempo total de 29 horas e 20 minutos e resultou em 1643 assuntos que estão listados no anexo **??**.

4.1.2 Segunda fase de processamento: queries e narrativas da coleção pública

A segunda fase de processamento da análise de assuntos ocorreu sobre a coleção de referência pública usada na trilha *Enterprise* da *Text Retrieval Conference* (TREC) até o ano de 2008. Trata-se de uma coleção de 370715 páginas *Web* públicas da empresa CSIRO; além de um conjunto de 77 narrativas e 77 *queries* disponibilizadas por funcionários da CSIRO como as questões mais frequentemente respondidas por funcionários da CSIRO para clientes externos (BAILEY et al., 2007a). Apenas as narrativas e *queries* foram adotadas nesta segunda fase. As narrativas correspondem a uma questão de exemplo escrita por um cliente e enviada para funcionários da CSIRO. Para responder a questão do cliente, um funcionário deve realizar uma ou mais buscas no sistema de recuperação de informação da empresa. Essa busca é constituída por um exemplo de *query* que retorna documentos pertinentes para responder a questão.

As narrativas e as *queries* apresentam termos e assuntos que podem ser vistos como índices para 19650 documentos classificados como “altamente relevantes” para as necessidades de busca de seus clientes e funcionários, o que corresponde 5,3% de toda a coleção composta por 370715 documentos. Dessa forma, as narrativas e as queries foram usadas partindo do pressuposto de que elas são suficientes como índice para os termos e assuntos mais frequentes para clientes e especialistas de informação da empresa.

O mesmo profissional da primeira fase leu as 77 narrativas e *queries* em pares enquanto anotava os assuntos presentes. Nessa etapa, os assuntos foram anotados através do termo em linguagem natural que melhor o representava pelo emprego da terminologia adotada pelo cliente (no caso das narrativas), pela instituição (no caso das *queries*) ou por ambos.

Após a leitura de cada par narrativa-*query* e a anotação dos novos assuntos descobertos, foram aplicadas as técnicas de dissecação e desnudação para formação de assuntos adicionais.

Tendo em vista que as narrativas e *queries*, as quais compõem a amostra da segunda fase, apresentam um escopo mais limitado, nenhum assunto foi produzido por dissecação. Através do método de desnudação, pelo qual se produz cadeias de assuntos em que a profundidade aumenta e a extensão diminui a cada iteração do método, houve formação de assuntos. Um exemplo de formação de assuntos pelo método de desnudação nessa fase é a cadeia *Australia* $\supset$ *New South Wales* $\supset$ Sydney para a hierarquia espacial da cidade de Sydney. Ou seja, *Sydney* está contida em *New South Wales*, que está contida em *Australia*.

Os demais três métodos de formação de assuntos, laminação, agregação e sobreposição, não foram utilizados na amostra, uma vez que foi usada a linguagem natural presente

na amostra, inexistindo vocabulário controlado para controle terminológico.

Os assuntos descobertos na primeira e segunda fases de processamento foram categorizados e as duas coleções sofreram uma primeira comparação antes que a terceira fase de processamento acontecesse. Os procedimentos e o resultado da categorização encontra-se na seção 4.2; os procedimentos e o resultado da comparação entre as coleções encontra-se na seção 4.3; porém, a execução dos procedimentos e obtenção dos resultados da terceira fase, descritos na próxima seção 4.1.3, caso realizados imediatamente após as duas primeiras fases, não alterariam a comparação entre coleções. A análise e formação de assuntos de narrativas e *queries* da coleção pública consumiu o tempo total de 42 horas e 11 minutos e resultou em 241 assuntos que estão listados no anexo **??**.

4.1.3 Terceira fase de processamento: documentos da coleção pública

Na terceira fase de processamento foi feita a análise de assuntos sobre a coleção de referência pública usada na trilha *Enterprise* da *Text Retrieval Conference* (TREC) até o ano de 2008. Porém, nesta fase o interesse eram os documentos da coleção e não mais as narrativas e *queries*. Como a coleção é constituída por 370715 páginas *Web* públicas e sua leitura demandaria mais tempo que aquele possível para a pesquisa, uma amostra de apenas 50 documentos foi analisada intelectualmente.

Para constituir a amostra, foram selecionados os 19650 documentos considerados altamente relevantes para as 77 narrativas e *queries* processadas na segunda fase de processamento. Os documentos foram ordenados pela quantidade de narrativas e *queries* em que eram considerados altamente relevantes. Os 50 documentos mais frequentes em narrativas e *queries* (0,254% do total de documentos altamente relevantes) foram então selecionados para a terceira fase, sendo comuns a um mínimo de 25 narrativas e *queries*. Em média, os 50 documentos estão associados a 36,46 narrativas e *queries* com desvio padrão de 6,8756.

O mesmo profissional das duas primeiras fases leu os 50 documentos enquanto anotava os assuntos presentes. Nesta fase, os assuntos foram anotados através do termo em linguagem natural que melhor o representava pelo emprego da terminologia adotada pela instituição. Após a leitura de cada documento e a anotação dos novos assuntos descobertos, foram aplicadas as técnicas de dissecação e desnudação para formação de assuntos adicionais.

Através do método de dissecação são produzidos renques dentro de facetas. Um exemplo de formação de assuntos pela aplicação da técnica de dissecação é o renque New South Wales (NSW), Queensland (Qld), South Australia (SA), Tasmania (Tas), Victoria (Vic) e Western Australia (WA) de estados australianos.

Através do método de desnudação são identificados assuntos com grande profundidade a partir de assuntos mais gerais ou isolados.Um exemplo de formação de assuntos pelo método de desnudação, é a cadeia não exaustiva *Energy Technology* $\supset$ *Energy Centre* $\supset$ *National Solar Energy Centre*, de unidades organizacionais da divisão de pesquisa da CSIRO em tecnologia energética. Ou seja, *National Solar Energy Centre* está contida na unidade *Energy*

Centre, que está contida na diretoria de *Energy Technology*.

Os demais três métodos de formação de assuntos, laminação, agregação e sobreposição, não foram utilizados na amostra, uma vez que foi usada a linguagem natural presente na amostra, inexistindo vocabulário controlado para controle terminológico.

Os assuntos descobertos foram categorizados e as facetas resultantes foram comparadas com aquelas das duas fases anteriores, como é visto nas seções 4.2 e 4.3 respectivamente. A terceira fase de análise e formação de assuntos na coleção pública consumiu um tempo adicional de 27 horas e 51 minutos e resultou em 913 assuntos que estão listados no anexo **??**.

4.2 Categorização de assuntos através da análise facetada

Após o processamento de cada uma das três fases descritas na seção anterior, os assuntos reconhecidos pelo processo de análise foram categorizados em uma estrutura classificatória com três níveis. O primeiro nível representa as categorias fundamentais, o segundo nível representa as facetas pelas quais cada categoria é dividida, enquanto o terceiro nível constitui as subfacetas pelas quais cada faceta é dividida.

Categorias, facetas e subfacetas emergiram da análise facetada pela necessidade de categorizar os assuntos reconhecidos nas amostras. Isto é, a estrutura não foi definida arbitrariamente antes de o processo de categorização ocorrer. Os nomes das categorias, facetas e subfacetas não servem para atribuir significado. Por outro lado, o significado de cada uma delas pode ser deduzido facilmente pela posição em que cada uma ocupa na estrutura classificatória facetada.

As categorias, facetas e subfacetas consolidadas a partir das três amostras são:

- Categoria **Associação** – onde residem os assuntos relacionados às instituições associadas a instituição proprietária da coleção – dividida em
    1. Aquisição
        a) Nome de empresa
    2. Atendimento
        a) Nome de recurso
    3. Cliente
        a) Nome de unidade
    4. Comunicação
        a) Nome de veículo
        b) Tipo de veículo
    5. Concorrência
        a) Apelido de empresa
        b) Nome de empresa
        c) Nome de serviço
    6. Financiador
        a) Nome de empresa
    7. Fornecedor
        a) Nome de empresa
        b) Nome de serviço
        c) Fusão
    8. Grupo
        a) Nome de empresa
    9. Influência
        a) Apelido de pessoa
        b) Nome de empresa

   c) Nome de pessoa
   d) Nome de unidade
   e) Papel de pessoa
   f) Poder público
10. Parceria
   a) Apelido de empresa
   b) Modalidade
   c) Nome de empresa
   d) Nome de pessoa
   e) Site de empresa
   f) Tipo de parceiro
11. Parte
   a) Função
   b) Nome de empresa
   c) Profissional
12. Sindicato
   a) Nome de empresa

- Categoria **Comunicação** – onde residem os assuntos relacionados aos canais de comunicação mobilizados na instituição – dividida em
   1. Correspondência
   2. Email
   3. Internet
   4. Rádio
   5. Telefone
   6. Televisão
- Categoria **Conhecimento** – onde residem os assuntos relacionados às áreas de conhecimento e profissionais – dividida em
   1. Área
- Categoria **Documento** – onde residem os assuntos relacionados aos documentos, seu ciclo de vida, suas características físicas e metadados – dividida em
   1. Acervo
      a) Banco de dados
      b) e-book
      c) Livro
      d) Material
   2. Anais
   3. Artigo
   4. Assunto
   5. Atualização
   6. Autorização
   7. Avaliação
   8. Blog
   9. Coleção
   10. Conjunto
   11. Contrato
      a) Alteração
   12. Entrega
      a) Data
      b) Meio
   13. Fluxo
   14. Folheto
   15. Formato de documento
   16. Formulário
   17. Gênero de documento
      a) Ata
      b) Carta
      c) Certificado
      d) Declaração
      e) Edital
      f) Extrato
      g) Fluxograma
      h) Laudo
      i) Notícia
      j) Parecer
      k) Planilha
      l) Portaria
      m) Projeto

    n) Questionário
    o) Resolução
18. Identificação
19. Idioma
20. Imagem
21. Legislação
22. Lista
23. Livreto
24. Livro
25. Manual
26. Mapa
27. Marca
28. Modelo
29. Norma
30. Palavra-chave
31. Parte
32. Podcast
33. Portifólio
34. Preservação
35. Produção
36. Regulamento
37. Relatório
38. Réplica
39. Requerimento
    a) Assunto
    b) Situação
40. Roteiro
41. Site
    a) Endereço
42. Suporte
43. Tamanho
44. Versão
45. Vídeo

- Categoria **Economia** – onde residem assuntos relacionados às características das atividades econômicas e da moeda – dividida em
    1. Arranjo produtivo
    2. Atividade econômica
        a) Educação
    3. Moeda
    4. Porte
- Categoria **Espaço** – onde residem assuntos georreferenciáveis – dividida em
    1. Cidade
        a) Apelido de cidade
        b) Bairro
        c) Endereço
        d) Equipamento urbano
        e) Logradouro
        f) Região
        g) Região urbana
    2. Continente
    3. Distrito federal
    4. Estado
        a) Região estadual
    5. Medida
    6. País
        a) Região
        b) Rodovia
    7. Região
        a) Bloco econômico
- Categoria **Instituição** – onde residem assuntos relacionados à empresa e às unidades organizacionais da empresa – dividida em
    1. Apelido de empresa
    2. Atualização

3. Cultura
4. Nome de empresa
5. Tipo de instituição
6. Unidade
    a) Papel
7. Visão

- Categoria **Operação** – onde residem assuntos relacionados às atividades de produção e de gerenciamento dos processos organizacionais da empresa – dividida em

    1. Atendimento
        a) Área de atendimento
        b) Capacidade
        c) Desempenho
        d) Equipe
        e) Expansão
        f) Frequência
        g) Interrupção
        h) Método
        i) Nome de evento
        j) Pós-venda
        k) Produto
        l) Projeto
        m) Tipo de evento
    2. Cobrança
    3. Compra
        a) Pagamento
    4. Controle
        a) Alocação
        b) Análise
        c) Apresentação
        d) Avaliação externa
        e) Decisão
        f) Desempenho
        g) Discussão
        h) Equipe
        i) Frequência
        j) Pendência
        k) Reunião
        l) Sanção
        m) Seleção
    5. Desenvolvimento
        a) Avaliação
        b) Equipe
        c) Papel
        d) Produto
        e) Projeto
    6. Divulgação
        a) Instrumento
        b) Meio
        c) Nome de evento
        d) Público-alvo
        e) Tipo de evento
        f) Venda
    7. Estoque
    8. Financiamento
        a) Captação
        b) Modalidade
        c) Seleção
        d) Informática
        e) Nome de sistema
    9. Manutenção
        a) Equipe
    10. Orçamento
        a) Alocação
    11. Pessoal
        a) Admissão
        b) Alocação
        c) Avaliação
        d) Capacitação
        e) Competência
        f) Demissão
        g) Equipe
        h) Inventário

i) Licença
j) Nome de evento
k) Plano de saúde
l) Progressão
m) Remuneração
n) Seleção
o) Transferência
p) Vaga

12. Produto
13. Segurança
14. Situação
15. Suporte à operação
    a) Apelido de recurso
    b) Infraestrutura
    c) Nome de recurso
16. Transporte
    a) Equipamento
    b) Evidência
17. Venda
    a) Produto

- Categoria **Patrimônio** – onde residem assuntos relacionados aos bens móveis e imóveis – dividida em
  1. Atualização
  2. Depreciação
  3. Equipamento
  4. Imóvel
     a) Construção
     b) Equipamento
     c) Infraestrutura
     d) Projeto
  5. Licença de software
  6. Participação societária
- Categoria **Pessoal** – onde residem assuntos relacionados às pessoas naturais, internas e externas à instituição – dividida em
  1. Cliente
     a) Conjunto
     b) Idade
     c) Papel
     d) Responsável
     e) Situação
     f) Tempo de vínculo
  2. Comunidade
  3. Desenvolvimento
  4. Externo
     a) Apelido
     b) Nome
     c) Papel
     d) Profissão
     e) Situação
  5. Filiação
  6. Fornecedor
     a) Prestador de serviço
  7. Grupo
     a) Papel
  8. Identificação
  9. Profissional
     a) Competência
     b) Conjunto
     c) Email
     d) Equipe
     e) Formação
     f) Função
     g) Lotação
     h) Modalidade
     i) Modalidade de contratação
     j) Nome
     k) Origem
     l) Papel
     m) Profissão
     n) Remuneração
     o) Sobrenome

   p) Status
   q) Tempo de vínculo
   r) Título acadêmico
10. Sexo

- Categoria **Procedimento** – onde residem assuntos relacionados a tarefas desenvolvidas internamente na instituição – não dividida em facetas
- Categoria **Tempo** – onde residem assuntos de referência temporal – dividida em

1. Alocação
2. Ano
3. Ano especial
4. Atualização
5. Calendário
6. Cronograma
7. Data
8. Década
9. Evento
   a) Tipo de evento
10. Fuso horário
11. Hora
12. Hora especial
13. Horário
14. Mês
15. Período
   a) Intervalo
   b) Intervalo em anos
   c) Intervalo em dias
   d) Intervalo em horas
   e) Intervalo em meses
16. Projeção
17. Século
18. Semana
19. Valor

Então, os assuntos passaram pelo processo de categorização da análise facetada e as categorias encontradas podem ser vistas como um meio de comparação entre as coleções públicas e privadas. A análise facetada ocorreu sobre cada um dos assuntos identificados em cada fase do processamento de assuntos, com conferência e correções ao longo do processo. Após concluir a análise facetada sobre os três conjuntos de assuntos, uma nova iteração de correções e ajustes ocorreu com o objetivo de garantir consistência à categorização, entre os assuntos e as coleções.

O processo de síntese da técnica de análise facetada, quando assuntos compostos são formados pela combinação de conceitos, não foi implementado nesta tese. Isso se justifica pela importância de se reconhecer assuntos e termos relevantes para autores e usuários de cada coleção, ao invés de se tentar enumerar as quase infinitas possibilidades de assuntos que podem ocorrer nas coleções.

A formação das categorias respeitou os nove princípios do plano das ideias compilados por Spiteri (1998) e discutidos na seção 3.2.2.1. O reconhecimento e a atribuição de significado à terminologia respeitou os dois princípios do plano verbal compilados por Spiteri (1998) e ainda respeitou o cânone da reticência de Ranganathan (1967), explicados e discutidos na seção 3.2.2.2. Os princípios do plano notacional não foram considerados nesta tese, uma vez que não foi produzido um sistema notacional para o sistema de classificação. Em outras palavras, a principal linguagem de trabalho usada pelo profissional da área e pelo usuário da

informação foi a própria linguagem natural presente nos documentos.

Um exemplo de resultado da categorização é o termo “bluetongue disease”, categorizado como Operação-Desenvolvimento-Produto. Outros termos foram categorizados através de dois níveis da estrutura de classificação. É o caso do termo “new south wales”, categorizado como Espaço-Estado. Também há termos categorizados através de apenas um nível, tal como o termo “identification” categorizado como Procedimento. Adicionalmente, um termo pode ser categorizado diferentemente entre as coleções e documentos, uma vez que a atenção é voltada para o significado e para a categoria do significado associados com o termo. No entanto, o método de análise facetada tentou ajustar sua atenção para todo o domínio constituído por ambas as coleções ao invés de restringir-se apenas ao escopo de um único e isolado documento, de apenas uma das coleções. Em outras palavras, a análise facetada não foi usada para categorizar documentos ou termos, mas para representar os assuntos emergidos dos documentos através de termos da linguagem natural.

Ao final, a avaliação da estrutura classificatória de três níveis ocorreu sobre o resultado consolidado das duas coleções. As categorias, facetas e subfacetas resultantes são analisadas na próxima seção 4.3.

4.3 Avaliação de categorias, facetas e subfacetas

A avaliação da estrutura classificatória e a comparação entre as duas coleções deram-se em cada um dos níveis da classificação, sendo que o primeiro nível está associado às categorias fundamentais, o segundo nível está associado às suas facetas e o terceiro nível está associado às subfacetas. Os três níveis são tratados nas subseções 4.3.1, 4.3.2 e 4.3.3, respectivamente.

4.3.1 Categorias

Doze categorias emergiram da análise facetada e agrupam os assuntos resultantes dos documentos da coleção particular, das narrativas e *queries* da coleção pública e dos documentos da coleção pública. As categorias são Pessoal, Associação, Instituição, Documento, Comunicação, Economia, Conhecimento, Operação, Procedimento, Espaço, Patrimônio e Tempo. A figura **??** ilustra a frequência com a qual assuntos foram categorizados em cada uma das categorias do primeiro nível da estrutura classificatória, onde a coleção pública ainda encontra-se dividida entre as narrativas e *queries* (*queries* da CSIRO) e os documentos (documentos da CSIRO).

Em ordem decrescente, as categorias Operação, Documento e Pessoal são as categorias com mais assuntos em qualquer das coleções. Isso indica que o conteúdo corporativo frequentemente apresenta características próprias para diferentes etapas da sua atividade econômica, para diferentes gêneros de documento e para diferentes indivíduos que atuam na empresa. Em outras palavras, é possível afirmar que o conteúdo corporativo orbita em torno das suas operações; se manifesta em gêneros textuais e de documentos, que apesar de poucos são explicitados por sua função de dar forma ao conteúdo; e são produzidos por e

endereçados a indivíduos, ou ainda podem usar da autoridade de indivíduos para ter aumentada sua difusão. No entanto, a categoria Operação isoladamente agrupa entre 30% e 50% de todos os assuntos que ocorrem nas coleções, as categorias Documento e Pessoal, isoladamente, agrupam um máximo de 15% de todos os assuntos que ocorrem em cada coleção e as demais categorias em média agrupam 10% de todos os assuntos.

A figura **??** também agrupa as 12 categorias resultantes pela sua natureza tal como Entidades sociais, Mensagens, Disciplinas científicas, Processos de negócios, Espaço e Tempo. Por essa estrutura, o grupo de Processos de negócios, onde se encontra a categoria Operação, é aquele com o maior número de assuntos, mesmo sem o suporte da categoria Procedimentos. Entidades sociais, o segundo maior grupo, agrupa as categorias Pessoal (humanos), Instituição (pessoas jurídicas) e Associação (relações entre humanos, pessoas jurídicas e ambos). Em seguida, Mensagens posiciona-se como terceiro maior grupo principalmente por agrupar os assuntos da categoria Documentos, mas equipara-se ao grupo de Espaço. Espaço agrupa duas categorias, Espaço e Patrimônio, sendo que a primeira está ligada a nomes geográficos comuns ao ambiente corporativo e a segunda está ligada a equipamentos corporativos que podem ser georreferenciados. O quarto maior grupo é o Tempo, constituído por uma categoria com o mesmo nome, o que reflete a importância do tempo para as mensagens corporativas, dando a elas contexto temporal e validade limitada. Finalmente, Disciplinas científicas agrupa as categorias de Economia (atividades econômicas, principalmente) e Conhecimento (profissões e campos de estudo, principalmente). O último grupo demonstra o quanto esses assuntos são escassos nas mensagens corporativas, possivelmente porque as profissões e áreas de conhecimento sejam conhecidas *a priori* pelos receptores das mensagens corporativas e portanto não precisam ser marcadas explicitamente no conteúdo das mensagens.

Os seis agrupamentos de categorias são muito semelhantes às cinco categorias fundamentais de Ranganathan, sendo o grupo Entidades sociais compatível com a categoria Personalidade; os grupos Mensagens e Disciplinas científicas compatíveis com a categoria Matéria; o grupo Processos de negócios compatível com a categoria Energia; e os grupos Espaço e Tempo compatíveis com as categorias fundamentais de mesmo nome. Associando-os dessa forma, a categoria fundamental energia constitui aquela com maior frequência nas coleções investigadas (com cerca de 50% dos assuntos existentes), enquanto as outras categorias fundamentais apresentam igual frequência (com cerca de 10% dos assuntos existentes, cada uma).

As observações sobre categorias e grupos de categorias são comuns às duas coleções. As figuras **??** e **??** demonstram graficamente a correlação entre o resultado da categorização entre documentos da coleção particular e pública, e entre o resultado da categorização entre documentos da coleção particular e narrativas e *queries* da coleção pública.

Adicionalmente, pela figura **??** pode ser observado que tanto as mensagens corporativas quanto as buscas construídas por profissionais de informação, passando pelas mensagens de clientes, apresentam a mesma distribuição de assuntos e facetas. Isto é, em uma busca, os usuários informacionais tendem a incluir assuntos associados às facetas identificadas com alta

correlação positiva se comparados aos assuntos presentes nos documentos de interesse. A figura **??** compara documentos da coleção pública com narrativas e *queries* da coleção pública, enquanto a figura **??** compara documentos da coleção particular com narrativas e *queries* da coleção pública.

Apesar de as coleções apresentarem semelhanças no nível mais alto da estrutura classificatória, as diferenças surgem e aumentam na medida em que níveis mais precisos de classificação são mobilizados. A próxima seção 4.3.2 avança um nível na hierarquia da classificação e trata das facetas pelas quais cada categoria foi dividida.

4.3.2 Facetas de categorias

Ao avançar para o segundo nível da estrutura de classificação, as 12 categorias fundamentais são divididas em 148 facetas diferentes. Entre os assuntos da coleção particular, do conjunto de narrativas e *queries*, e da amostra documentos da coleção pública, 97,56%, 95,02% e 98,13% dos assuntos são classificados pelo menos até o nível de facetas.

Em ordem alfabética, Associação é a primeira categoria, dividida em 14 facetas. Dentre elas, a faceta Parceria é a mais frequente e reúne assuntos relacionados a entidades que tenham relação de cooperação com a empresa original. Outras facetas populares em ordem alfabética são Comunicação, Concorrência, Fornecedor e Parte, relacionadas principalmente a veículos externos de difusão de informação, entidades concorrentes, entidades fornecedoras de produtos e serviços, e entidades onde a empresa possui participação societária, respectivamente. No entanto, as coleções não apresentam correlação na distribuição de facetas dentro da categoria Associação, o que sugere que as coleções tratam de interesses institucionais diferentes. Se setores compatíveis de duas ou mais organizações possuem correlação é uma questão em aberto.

A segunda categoria, Comunicação, é dividida em seis facetas que representam meios de comunicação como Correspondência, Email, Internet, Rádio, Telefone e Televisão. Embora assuntos associados a esses meios de comunicação sejam comuns, especialmente endereços de email e números de telefone, não há correlação entre as coleções.

A terceira categoria, Conhecimento, ganhou uma única faceta pela aplicação da técnica de desnudação. Assim, Conhecimento agrupou apenas o assunto de mais alto-nível ciência e todas as áreas do conhecimento e profissionais foram agrupadas na faceta Área. Essa estrutura é acidental e não se justifica. Uma alternativa viável é tomar a categoria Conhecimento como áreas profissionais e do conhecimento e não dividí-la em facetas, mantendo um único nível. O número de assuntos na classe Conhecimento-Área é menor que 0,07% em todas as coleções, o que torna opcional qualquer alteração em sua estrutura.

A quarta categoria, Documento, conta com 45 facetas. Não há correlação entre as coleções dentro da categoria Documento, porém as facetas mais populares em cada coleção mostram-se úteis. A faceta mais comum na coleção particular é Gênero de documento, onde estão gêneros textuais e tipos de documentos que se mostraram numerosos. Na coleção pública, por sua vez, gêneros textuais e tipos de documentos não são declarados no conteúdo de

documentos, embora algumas diferenças significativas entre documentos possam ser notadas. Na coleção pública, muitos assuntos são classificados dentro da faceta Portifólio por representarem um pequeno guia (*hub*) que permite a navegação para mais informação supostamente de interesse ao usuário do portifólio.

Economia é a quinta categoria e foi dividida em quatro facetas. A faceta Arranjo produtivo tem elevado potencial de georreferenciamento e só foi usada na coleção particular. A faceta Atividade econômica é mais comum em todas as coleções, uma vez que atividades econômicas e profissionais são acomodadas comumente nessa faceta. As facetas Moeda e Porte acomodam poucos assuntos normalmente associados ao câmbio e à classificação de tamanho de empresas mais comum para a região onde a coleção se originou. Não há correlação entre as coleções na categoria Economia.

Espaço é a sexta categoria e foi dividida em sete facetas. As facetas são resultado da desnudação baseada na hierarquia geográfica, onde temos as facetas Cidade, Continente, Distrito federal, Estado, Medida, País e Região. Apenas a faceta Medida não merece participar do renque por agrupar unidades de medida espacial ao invés de nomes de lugar. As facetas mais comuns são Cidade, Estado e País, justificado pela forma como as pessoas se referem a espaços na linguagem cotidiana quando não precisam de grande precisão geográfica. Exatamente por esse motivo, a distribuição de assuntos em facetas da categoria Espaço apresenta alta correlação positiva entre as duas coleções.

A sétima categoria, Instituição, também foi dividida em sete facetas. As facetas confundem-se com características organizacionais bem conhecidas, como Apelido de empresa, Atualização, Cultura, Nome de empresa, Tipo de instituição, Unidade e Visão. As coleções não apresentam correlação estatística, apesar do coeficiente indicar o contrário. Entre documentos, uma correlação positiva próxima de 1 entre as amostras de documento apenas sugere que ambas as empresas contam com muitas unidades organizacionais classificadas na faceta Unidade, além de possuírem missão, visão, nome empresarial e outros atributos que realmente são características populares entre empresas. Com isso, foram as grandes estruturas organizacionais, refletidas na faceta Unidade, que determinaram a similaridade entre as duas coleções.

A estrutura organizacional também interferiu em Operação. A oitava categoria foi dividida em 18 facetas, sendo que algumas facetas são atividades de unidades da instituição (Instituição-Unidade), enquanto outras são atividades de setores administrativos não presentes no organograma ou podem ser atividades administrativas presentes em diversos locais do organograma. As facetas são Atendimento, Cobrança, Compra, Controle, Desenvolvimento, Divulgação, Estoque, Financiamento, Informática, Manutenção, Orçamento, Pessoal, Produto, Segurança, Situação, Suporte à operação, Transporte e Venda. Não há correlação estatística entre as duas coleções, o que sugere que o público-alvo e/ou propósito dos documentos das coleções sejam diferentes. A hipótese de que documentos com público-alvo e propósito compatíveis apresentem alta correlação estatística merece ser verificada, algo que requer um estudo adicional fora do escopo da presente pesquisa. As facetas que em média apresentam o maior número de assuntos são Desenvolvimento, Atendimento, Pessoal e Controle, que su-

gerem a complexidade da comunicação no desenvolvimento de novos produtos e serviços, no atendimento de clientes, na administração de pessoas, e nos processos decisórios, respectivamente.

Patrimônio é a nona categoria e foi dividida em seis facetas. As facetas são Atualização, Depreciação, Equipamento, Imóvel, Licença de software e Participação societária e não apresentam correlação estatística entre as duas coleções. De fato, bens móveis e imóveis são explicitados principalmente em função da atividade econômica, uma diferença importante entre as duas empresas investigadas. Por outro lado, a distribuição de assuntos entre essas facetas pode ajudar a classificar empresas do mesmo porte e da mesma atividade econômica. As facetas que apresentam maior número de assuntos são Imóvel e Equipamento.

A décima categoria, Pessoal, foi dividida em dez facetas. As facetas são Cliente, Comunidade, Desenvolvimento, Externo, Filiação, Fornecedor, Grupo, Identificação, Profissional e Sexo. Foi registrada alta correlação positiva na distribuição de assuntos entre as facetas da categoria Pessoal, especialmente porque indivíduos têm sido representados de forma muito semelhante em documentos das coleções investigadas. As facetas com maior média de assuntos são Profissional, Externo e Cliente que acomodam indivíduos que trabalham na empresa, que colaboram com ou influenciam a empresa, e que fazem uso de serviços da empresa, respectivamente.

A décima primeira categoria, Procedimento, não foi dividida em facetas ou em subfacetas. No entanto, é possível que a ausência de documentos normativos das empresas explique o fenômeno. Sua ausência sugere que as empresas não têm interesse de dar ampla publicidade a certos documentos, mantendo-os restritos exclusivamente aos funcionários para quem os documentos se destinam.

Por último, as facetas da categoria Tempo são 19, sendo que as facetas Período, Ano, Calendário e Cronograma apresentaram mais assuntos. Não há correlação na distribuição dos assuntos de facetas da categoria Tempo entre as coleções.

Apesar de as coleções apresentarem semelhanças no nível mais alto da estrutura classificatória, suas especificidades mostram-se óbvias no segundo nível de classificação. A próxima seção 4.3.3 avança um nível na hierarquia da classificação e trata das subfacetas pelas quais cada faceta foi dividida.

4.3.3 Subfacetas de facetas

Ao avançar para o terceiro nível da estrutura de classificação, as 148 facetas discutidas na seção anterior foram divididas em 311 subfacetas. Entre os assuntos da coleção particular, do conjunto de narrativas e *queries*, e da amostra documentos da coleção pública, 63,60%, 58,09% e 65,61% dos assuntos são classificados pelo menos até o nível de subfacetas.

Porém, comparando as duas coleções, a distribuição de assuntos em subfacetas, no terceiro nível, não apresenta qualquer similaridade. As especificidades de cada coleção tornaram-se evidentes e tentativas de acomodar assuntos de uma coleção em subfacetas comuns às duas coleções mostraram-se ineficazes.

4.4 Discussões

A generalização do domínio corporativo apresentada é preliminar e constitui apenas uma proposta. Sua primeira restrição está na pequena razão entre o número de empresas pesquisadas, apenas duas, e o grande número de empresas existentes no mundo. Além de muito tempo necessário para analisar vários documentos e várias empresas, reunir informação de um grande número de empresas é especialmente desafiador uma vez que as empresas normalmente não revelam sua informação. Afinal, revelar informação exporia clientes, funcionários, parceiros comerciais e planos futuros (BAILEY et al., 2007a). A segunda restrição está na própria natureza do domínio corporativo que reúne empresas de setores, atividades, idiomas, culturas, tamanhos e missões tão diversos. Assim, mesmo uma amostra composta de um número elevado de empresas dificilmente bastaria para representar todo o domínio corporativo e para suportar o projeto de sistemas de recuperação de informação melhores para todas as demais empresas (HALEVY et al., 2005). Finalmente, uma terceira restrição está na profundidade limitada da análise do domínio empreendida. Uma vez que o empirismo e métodos estatísticos não remetem ao porquê e ao limite temporal da existência de certos padrões, sua interpretação requer análise mais aprofundada, histórica e racional da natureza, do propósito e do uso da informação do domínio corporativo (HJØRLAND, 2002), algo que mobilizaria muito mais recursos que o faz um único trabalho exploratório, por um único pesquisador.

Mesmo com limitações, a análise preliminar do domínio baseada na análise facetada foi útil por permitir um desenvolvimento gradual, incremental e flexível de uma classificação facetada para o domínio corporativo. Seu resultado foi uma estrutura classificatória suficiente para subsidiar a construção de um sistema de recuperação de informação corporativo que seja comum às duas empresas analisadas, mas também constitui um esquema de classificação mais facilmente ajustável às características da informação presentes em outras empresas.

O tempo de processamento da primeira fase, de documentos da coleção particular, mostrou-se muito inferior ao tempo de processamento da segunda e da terceira fases, ambas da coleção pública, mesmo para um número muito maior de ítens processados. A principal justificativa para a diferença de tempo é o conhecimento prévio do único classificador empregado sobre a empresa da coleção particular. A coleção pública requereu um tempo adicional para estudar a empresa, seus processos, a terminologia adotada, o território australiano e documentos complementares sobre o setor de atuação da empresa.

O número de documentos processados também difere entre as duas coleções. A coleção particular apresenta um menor número de documentos se comparada à coleção pública, embora a coleção particular pareça apresentar uma maior diversidade de tipos de documentos e representar um maior número de unidades organizacionais. Por outro lado, a coleção particular exigiu a leitura de mais documentos que a coleção pública, uma vez que a última contou com um índice construído previamente por terceiros que se mostrou adequado para o objetivo de compará-las.

Ambas as coleções também diferem em idioma, unidades organizacionais, atividades e público-alvo. Apesar dessas diferenças, os documentos processados são os mais frequentemente citados em ambas as coleções e há uma grande compatibilidade entre as facetas

mobilizadas para classificar os assuntos de cada uma.

Antes de comparar ambas as coleções, é preciso discutir a compatibilidade entre o conjunto de narrativas e *queries* e a amostra de documentos da coleção pública. Para isso, uma amostra de 50 documentos da coleção pública serviu para validar as narrativas e *queries* como índices para os documentos. A figura **??** demonstra graficamente a correlação estatística entre os dois conjuntos e a alta correlação positiva sugere que seus usuários buscam e escrevem documentos mobilizando as mesmas facetas. Assim, as narrativas e *queries* constituem um índice para os documentos públicos na medida em que são constituídos por termos significativos para uma necessidade informacional e uma lista de documentos que são adequados para aquela necessidade. Porém, não foram observadas diferenças na utilidade de cada uma quando comparadas isoladamente com a coleção particular.

Por esse motivo, a comparação entre documentos da coleção particular e narrativas e *queries* da coleção pública foi equivalente àquela entre documentos da coleção particular e documentos da coleção pública. De fato, a utilidade de ambos os conjuntos da coleção pública é a mesma e qualquer um deles pode ser usado para esse estudo sem quase qualquer diferença, como ilustrado pelas figuras **??** e **??**.

Enquanto a figura **??** compara as narrativas e *queries* da coleção pública com os documentos da coleção particular, a figura **??** compara os documentos da coleção pública com documentos da coleção particular. Usando correlação de Spearman, elas apresentam um ρ muito próximo, sendo respectivamente $\rho = 0,8365, n = 12, p < 0,0006932$ e $\rho = 0,8881, n = 12, p < 0,00004583$. Embora a correlação não possa ser usada para fazer generalizações sobre todo o domínio corporativo, sua medida ajuda a explicitar similaridades e diferenças entre pares de repositórios corporativos. Adicionalmente, ambas as coleções, pública e particular, podem ser vistas como compatíveis e a alta correlação positiva demonstra que usuários diferentes têm necessidades informacionais similares com características compatíveis. As necessidades informacionais, expressadas por mensagens dos autores para os destinatários corporativos, são representadas através de um grupo de facetas para pessoas, instituições, tipos de documentos, processos, espaço e tempo. A existência das mesmas facetas tornam ambos os repositórios coleções corporativas válidas para testar sistemas de recuperação de informação, uma vez que a coleção particular se mostra válida por garantia de usuário.

Como a categoria Operação isoladamente agrupa entre 30% e 50% de todos os assuntos que ocorrem nas coleções, isso explica o efeito positivo que estruturas classificatórias baseadas em atividades de negócio causam na classificação, recuperação e *ranking* de documentos corporativos. As categorias Documento e Pessoal, isoladamente, agrupam um máximo de 15% de todos os assuntos que ocorrem em cada coleção. As demais categorias em média agrupam 10% de todos os assuntos. Essa distribuição desigual parece ter sido a motivação para o desenvolvimento de modelos de recuperação de informação corporativa baseados exclusivamente nas categorias e facetas mais populares. Por outro lado, percebe-se um grande potencial de aumento de precisão dos modelos se forem consideradas outras facetas.

Isso é reforçado pela existência dos seis grupos pelas quais as 12 categorias são agrupadas, como apresentado na figura **??**. Os seis agrupamentos de categorias são muito

semelhantes às cinco categorias fundamentais de Ranganathan, sendo o grupo Entidades sociais compatível com a categoria Personalidade; os grupos Mensagens e Disciplinas científicas compatíveis com a categoria Matéria; o grupo Processos de negócios compatível com a categoria Energia; e os grupos Espaço e Tempo compatíveis com as categorias fundamentais de mesmo nome. Associando-os dessa forma a categoria fundamental energia constitui aquela com maior significado nas coleções investigadas (com cerca de 50% dos assuntos existentes), enquanto as outras categorias fundamentais apresentam igual significado, com cerca de 10% dos assuntos existentes, cada uma.

O postulado de Categorias Fundamentais de Ranganathan (1967) em conjunto com a pequena amostra usada nesse estudo constituem evidência anedótica para uma generalização do domínio corporativo, o que requer mais estudos tomando outras coleções corporativas. Porém, ao observar a compatibilidade entre as categorias propostas e as categorias fundamentais de Ranganathan não deseja-se estabelecer nenhuma falácia genética. De fato, nenhuma conclusão para o domínio pode ser feita a partir dessa observação sobre as coleções.

Mesmo assim, as categorias comuns às duas coleções são úteis para melhorar o modelo de recuperação de informação corporativa, tanto para as empresas a que pertencem as coleções quanto para aquelas em que sua informação apresente as mesmas facetas com distribuição compatível. As categorias do primeiro nível, por exemplo, podem produzir um impacto no modelo de recuperação de informação corporativa das duas empresas, sem distinção. As facetas e subfacetas, de segundo e terceiro nível, também podem produzir impactos no modelo de recuperação, porém de forma diferente em cada uma das coleções.

Por outro lado, o conjunto de categorias, facetas e subfacetas produzido não deve ser visto como o único ou o melhor. Todo o processamento de documentos foi realizado por um único profissional da área, algo que favorece a consistência de categorização entre coleções e assuntos, mas a utilidade da classificação resultante ainda deve ser avaliada por usuários da coleção particular ou por meio de garantia literária no caso da coleção pública. No entanto, embora outras soluções sejam possíveis pela adoção de outros profissionais-indexadores, não importa ao propósito deste trabalho avaliar a consistência inter-indexadores e nem mesmo produzir uma indexação, classificação e recuperação com o maior desempenho possível para cada coleção.

O capítulo 5 apresenta a validação de um modelo de recuperação de informação baseado na classificação facetada desenvolvida no presente capítulo. Para a validação foi implementado um protótipo funcional de sistema de recuperação de informação corporativa sobre a coleção pública. Também, foram avaliadas as expressões de busca propostas por usuários reais da coleção particular. Ambos, o protótipo e o conjunto de expressões de busca são detalhados e servem como resultados empíricos para validação da utilidade e da eficiência do modelo facetado de representação.

5 Recuperação automatizada da informação corporativa e facetada

6 Conclusão

“As palavras fogem quando precisamos delas e sobram quando não pretendemos usá-las.”
Carlos Drummond de Andrade

Esta pesquisa justifica-se pela necessidade de se implementar sistemas automáticos de recuperação de informação que deem suporte adequado à informação corporativa e às tarefas dos usuários corporativos. Para isso, se buscou uma compreensão mais geral da informação corporativa que beneficie sua evolução contínua e não a limite a um cenário de uso excessivamente reduzido. Para isso, foi realizada uma análise do domínio corporativo com o objetivo de propor um conjunto de características da informação corporativa.

Assim, este trabalho empreendeu uma análise preliminar de domínio pela aplicação da técnica de análise facetada sobre informação corporativa. Foram identificadas características potencialmente comuns às coleções; e um protótipo de sistema de recuperação de informação corporativa e um conjunto de expressões de busca de usuários reais serviram para avaliar empiricamente e validar essas características identificadas.

A seção 6.1 apresenta os principais resultados alcançados, enquanto a seção 6.2 apresenta discussões e limitações da pesquisa. Finalmente, a seção 6.3 aponta algumas direções para trabalhos futuros.

6.1 Resultados

Três resultados foram obtidos neste trabalho. O primeiro resultado refere-se a um conjunto com 12 categorias que, como uma potencial representação da informação corporativa, pode servir como um modelo conceitual de longo prazo do domínio corporativo. O conjunto de categorias descobertas deve requerer revisões menos frequentes e suportar o desenvolvimento incremental de sistemas de recuperação de informação corporativa mais flexíveis e interoperáveis. O segundo resultado refere-se a um subconjunto das 12 categorias que é mobilizado especialmente por usuários no momento de elaborar expressões de busca com o objetivo de recuperar documentos de interesse. O terceiro resultado refere-se à validação de ambas as coleções corporativas como pertinentes para desenvolver e avaliar sistemas de recuperação de informação corporativa.

O primeiro resultado teve sua origem na análise facetada empreendida sobre os dois exemplares do domínio corporativo. Ele corresponde à identificação de 12 categorias comuns às duas coleções corporativas. A distribuição de assuntos de ambas as empresas dentro das categorias apresentou alta correlação positiva. Isso sugere que autores e leitores, de ambas as empresas, embora necessitem de assuntos diferentes, mobilizam assuntos das mesmas categorias e facetas.

Além das categorias identificadas, foram também avaliadas as facetas e subfacetas. Embora úteis para cada coleção, a distribuição de assuntos em facetas e subfacetas entre as coleções não apresenta indício de correlação. Isso sugere que características exclusivas de algumas empresas requerem facetas e subfacetas especiais que favorecem a comunicação de seus atores sociais. O volume de atores sociais contidos na empresa, a extensão geográfica atendida por seus serviços e produtos, as atividades e o setor econômico são exemplos de características exclusivas que parecem interferir na composição das mensagens corporativas.

Em avaliação, as categorias não figuraram entre as fontes de evidência mais comuns dos experimentos apresentados na trilha *Enterprise* da *Text Retrieval Conference*. Entretanto, a simples implementação de um protótipo que considerou parte das categorias aumentou o desempenho do recuperação de informação. O aumento do desempenho ocorreu mesmo sem fazer uso de repositórios externos à empresa e sem o suporte de metabuscadores, algumas das estratégias adotadas em outros trabalhos. Comparando os resultados atuais com aqueles encontrados na literatura, facetas espaciais e temporais mostraram-se especialmente úteis, uma vez que quase não têm sido exploradas em trabalhos interpretativos e constituíram as fontes de evidência com maior potencial de contribuição para a recuperação e o *ranking* da informação corporativa.

O segundo resultado teve sua origem na análise facetada empreendida sobre as expressões de busca de usuários da coleção particular. As expressões de busca foram elaboradas com o objetivo de recuperar documentos previamente apresentados pelo usuário. Assim, o método serviu para investigar quais categorias e facetas eram reconhecidas no documento e mobilizadas pelo usuário enquanto usavam um sistema de recuperação de informação hipotético. A distribuição de assuntos em oito categorias, compatível com a distribuição de assuntos naquelas categorias descobertas no domínio, evidencia que os usuários mobilizam as categorias corretas para elaborar as expressões de busca.

Por outro lado, o experimento sobre a coleção pública evidenciou que os usuários da coleção pública, apesar de também mobilizarem as categorias corretas, não conseguem escolher corretamente os termos espaciais e temporais. Isso pode ser explicado pelo desconhecimento da precisão geográfica e temporal com a qual o documento foi produzido e indexado. No entanto, o primeiro resultado, discutido anteriormente, foi suficiente para compatibilizar a granularidade espacial e temporal de indexação e de recuperação.

O terceiro resultado corresponde à validação cruzada da coleção de referência e da coleção particular. Coleções de referência usadas para avaliação de Cranfield normalmente são alvos de críticas sobre sua utilidade limitada para o desenvolvimento científico. Essa dificuldade também se apresenta para a coleção de referência adotada neste trabalho. Para enfrentar essa dificuldade, foi adotada também uma coleção particular. Uma vez que a coleção particular foi produzida por seus próprios usuários para ser usada no contexto de trabalho e tomada de decisão, ela pode ser vista como uma coleção suficiente para empreender novos estudos sobre informação corporativa. Por outro lado, a compatibilidade entre a distribuição de assuntos entre as categorias, facetas e subfacetas de ambas as coleções faz com que a coleção pública igualmente possa ser vista como uma coleção válida para estudos sobre

informação corporativa.

Adicionalmente, uma vez que não havia qualquer iniciativa de construção de coleção de documentos corporativos em língua portuguesa, a coleção particular constitui um produto útil para trabalhos futuros. Porém, a aplicação da nova coleção difere da aplicação da coleção de referência, sendo mais apropriada para estudos interpretativos e históricos. A coleção da CSIRO, principalmente por seu tamanho e método de criação, é mais adequada para estudos empíricos baseados em métodos estatísticos e carece de uma maior diversidade de informação no escopo intraorganizacional.

6.2 Considerações e limitações

O número reduzido de empresas e coleções estudadas apresenta limitações para uma análise mais aprofundada e para uma generalização do domínio corporativo. Na coleção de referência pública, os documentos são de um conjunto muito restritivo, apenas da *Web* pública da empresa, com necessidades de informação que poderiam ser atendidas pelo próprio cliente da empresa, através de um bom sistema de busca da *Web*. Na segunda coleção, particular, há um número bem menor de documentos, porém com uma maior diversidade temporal e de atividades. Em nenhuma das coleções há exaustividade, o que nos remete à impossibilidade de representar todo e qualquer fenômeno do domínio, embora provavelmente representem os fenômenos mais comuns de ambas as empresas.

Por outro lado, mesmo se for possível alguma generalização sobre as características do domínio corporativo, ela certamente não representa isoladamente toda a comunicação da qual as empresas precisam para realizar trabalho. Com isso, as características locais ou exclusivas de cada empresa continuam a desempenhar um papel importante na compreensão da comunicação dos seus diversos atores sociais.

Ao mesmo tempo, todas as características identificadas, das mais gerais até as mais específicas, parecem suportar a melhoria do desempenho de um sistema de recuperação de informação corporativa. Porém, as características mais específicas tendem a ser úteis apenas no contexto particular de uma organização, enquanto as características mais gerais provavelmente contribuem para empresas de todo o domínio corporativo.

A recuperação de informação da coleção pública foi avaliada empiricamente utilizando-se apenas das características mais gerais (as categorias). Essa avaliação, tratada no capítulo 5, embora destaque o valor das características comuns a ambas as empresas, requer cautela. De fato, não é possível generalizar o domínio corporativo a partir das duas empresas pesquisadas, assim como não é possível deduzir que todo o arquivo das empresas possui as mesmas características. A linguagem corporativa, como fenômeno social, está sujeita a um desenvolvimento contínuo e dinâmico, adaptando-se às necessidades das instituições e de seus atores sociais. Acompanhar continuamente essas mudanças parece ser a única saída para manter sistemas de recuperação de informação corporativa eficazes ao longo do tempo, adaptando-os para novas realidades de comunicação que são construídas ao longo do tempo.

Adicionalmente, a classificação de relevância de documentos para cada tarefa de

busca da coleção pública não é bem documentada. Como a qualidade dessa classificação não pode ser atestada neste trabalho, os resultados do *ranking* podem ser piores ou melhores que aqueles aferidos em outros trabalhos.

Relevância também não foi objeto de avaliação na coleção particular. A compatibilidade entre as expressões de busca e os documentos da coleção particular representa a uniformidade da comunicação corporativa, especificada tanto nos documentos quanto nas expressões de busca dos seus usuários, ao invés de provável relevância de documentos. A uniformidade da comunicação pode contribuir para recuperar e ordenar documentos baseando-se na sua utilidade para o contexto de trabalho. Para isso, é preciso explorar métodos mais adequados para quantificar e avaliar sua utilidade, percepção incompatível com aquela de relevância.

Resultados empíricos de outras trilhas da *Text Retrieval Conference* têm indicado que avanços tecnológicos sobre coleções influenciam menos o desempenho dos sistemas de recuperação de informação que avanços tecnológicos sobre facetas específicas. Como um exemplo, avanços em técnicas de identificação, desambiguação e recuperação de facetas sociais no domínio corporativo parecem repercutir positivamente em vários contextos ou várias coleções do domínio. Por outro lado, avanços na compreensão e no desempenho de uma coleção de uma só empresa, mesmo que contribua para o desempenho de um sistema de recuperação de informação aplicado àquela coleção, não contribui para o desempenho do mesmo sistema para qualquer outra empresa.

Porém, é preciso esclarecer que aperfeiçoamentos baseados em uma faceta do domínio corporativo, e.g. faceta social, não implica no mesmo ganho de desempenho em outros domínios. Portanto, é preciso garantir que sejam realizados estudos sobre facetas e domínios específicos, sem esperar que sistemas de recuperação de informação adequados para um domínio também sejam adequados para outros domínios aparentemente semelhantes. Isso justifica trabalhos futuros, usando as mesmas coleções e usando coleções corporativas adicionais, o que enfrenta a dificuldade permanente de construir coleções de documentos que sejam válidas sem que comprometam o sigilo de alguns dados corporativos.

Por outro lado, o método de comparação através de categorias, facetas e subfacetas mostrou-se valioso por conta da baixa exposição de dados e da alta comparabilidade, podendo servir a coleções diferentes e a diferentes pesquisadores. A técnica de análise facetada mostrou-se útil para descobrir características e comparar organizações sem expor em excesso seus ativos de informação, garantindo uma comparação simples e direta das categorias, facetas e subfacetas que constituem sua informação corporativa. Esse uso da análise facetada não foi observado na literatura. Adicionalmente, a classificação facetada resultante, além de útil para trabalhos futuros sobre informação corporativa, constitui um exemplar de classificação facetada construído fora da biblioteca e para fins não-didáticos. Esse último produto contribui para aperfeiçoar os guias metodológicos e estudos empíricos acerca da própria técnica de análise facetada (WILD; GIESS; MCMAHON, 2009; LA BARRE, 2010).

6.3 Trabalhos futuros

Os resultados deste trabalho sugerem algumas direções promissoras de trabalho futuro: a exploração de coleções corporativas adicionais; a exploração mais aprofundada de características da informação corporativa; novos estudos sobre o uso e o desempenho de sistemas de recuperação de informação; e o estudo de indexação automática da informação facetada.

Os resultados demonstram que as empresas apresentam muitas semelhanças e também muitas diferenças entre si. Portanto, a criação de mais coleções de referência é uma necessidade permanente. É preciso reunir amostras de dados de um número maior de organizações, com características que sejam úteis para trabalhos futuros, sem que as pesquisas se desenvolvam sobre apenas uma ou algumas poucas coleções de referência. Ao mesmo tempo em que novas coleções surgem, é importante que versões atualizadas também sejam produzidas de modo a permitir uma avaliação mais histórica do desenvolvimento da documentação, da linguagem corporativa e das necessidades dos usuários de uma dada organização.

Apenas dessa forma, se pode alcançar uma representação mais generalista do domínio ou pode ser identificada a impossibilidade de alcançá-la. Embora essa necessidade seja sugerida neste trabalho, reunir tantas empresas com dados abertos não é tarefa trivial. Uma solução intermediária passa pela caracterização da informação de várias empresas sem que os dados sejam livremente expostos. A análise de assuntos e a análise facetada atendem esses requisitos.

Também é importante diversificar as atividades e setores econômicos. Além disso, dentro do domínio corporativo, é preciso identificar as especificidades que idiomas, setores econômicos, atividades econômicas, tamanho da rede social corporativa, tipos de documentos e gêneros textuais provocam na informação corporativa.

A análise preliminar de domínio apontou que as facetas e subfacetas descobertas acomodam assuntos que são específicos de determinados contextos de uso. É o caso de facetas e subfacetas de elevada precisão geográfica, dentro da categoria Espaço. Em empresas com escopo geográfico urbano, são comuns os nomes de bairro, nomes de cidades vizinhas, nomes de capitais de estados vizinhos e nomes de empresas que correspondam a referências geográficas (agências bancárias, supermercados, hospitais, dentro outras). Ao contrário, em empresas com escopo geográfico nacional, são incomuns essas referências e tornam-se mais comuns os nomes de estados, nomes de cidades, nomes de empresas que não correspondem a referências geográficas (grandes empresas e multinacionais), e nomes de países.

Na perspectiva do uso da informação, é preciso estudar e experimentar novos modelos de interação que se beneficiem da classificação facetada da informação, além da busca por palavras-chaves. É o caso da navegação facetada, por exemplo, que passa a ser possível e mostra-se mais flexível para atender a uma maior diversidade de contextos de uso e necessidades de informação. Também é preciso experimentar novos modelos de ordenação de resultados (*ranking*) que permita um ajuste fino e personalizado de facetas potencialmente mais úteis para certos grupos de usuários. Nesses casos, é preciso realizar estudos sobre

estruturas de dados específicas para informação facetada e o seu papel no desempenho e na eficácia de sistemas de recuperação de informação.

Na avaliação do impacto da organização facetada na eficiência da recuperação e no *ranking* da informação corporativa, deve-se compará-la ao desempenho obtido através das estratégias de vocabulário controlado, de dados ligados externos (*linked data*), de busca em texto completo, dentre outras.

Adicionalmente, aplicações que têm se beneficiado menos de estudos interpretativos, como técnicas de *data mining*, *big data* sobre dados corporativos e sistemas de respostas automáticas, merecem ser observadas pela lente da teoria da análise facetada. Pelo grande volume de dados manipulados, análises intelectuais sobre essas coleções parecem proibitivas. Por outro lado, se amostras de dados da empresa são analisadas e apontam indícios de compatibilidade entre si, a modelagem e a representação de uma grande massa de dados podem se tornar mais significativas.

Como os experimentos deste trabalho beneficiaram-se da recuperação mais eficiente de entidades presentes no conteúdo dos documentos, é preciso estudar a implicação da organização facetada especialmente nos sistemas de respostas automáticas. Uma possível vantagem da organização facetada refere-se ao potencial de identificar características das entidades, com grande precisão, favorecendo a capacidade de responder a perguntas complexas elaboradas por seus usuários.

Por fim, dada uma classificação facetada adequada, a indexação de documentos corporativos continua a representar um grande desafio, tendo em vista o número crescente de documentos que as empresas produzem e mobilizam para dar suporte às suas operações. Portanto, é uma direção útil de trabalho futuro desenvolver técnicas de reconhecimento automático de termos e de associação dinâmica de termos às categorias e facetas identificadas neste trabalho.

O processo de reconhecimento de termos e sua associação a facetas normalmente dependerá de um processo de desambiguação sempre que não houver vocabulário controlado. Este trabalho encontrou referências à dados ligados (*linked data*) como suporte para a desambiguação de informação na *Web*. Que os dados ligados são úteis para a desambiguação e para a implementação de sistemas de informação parece óbvio, mas de que modo eles podem suportar a desambiguação de entidades na informação corporativa e qual sua eficácia em empresas menores é uma questão em aberto. No entanto, pelo volume da informação produzida, empresas de todos os tamanhos requerem sistemas de informação automatizados para gerenciar e recuperar eficientemente seus ativos de informação. Este trabalho demonstrou que a Ciência da Informação possui um papel fundamental para responder de que forma as empresas têm usado tais sistemas, e de que forma sistemas de recuperação de informação corporativos devem ser desenvolvidos.

Referências

ALBANI, A.; DIETZ, J. L.; ZAHA, J. M. Identifying business components on the basis of an enterprise ontology. In: KONSTANTAS, D. et al. (Ed.). *Interoperability of enterprise software and applications*. London, UK: Springer, 2006. p. 335–347. Citado na página 28.

ALBRECHTSEN, H. Subject analysis and indexing: from automated indexing to domain analysis. *Indexer*, Sheffield, England, v. 18, n. 4, p. 219–219, 1993. Citado na página 21.

ALENCAR, R. O.; DAVIS JÚNIOR, C. A.; GONÇALVES, M. A. Geographical classification of documents using evidence from wikipedia. In: WORKSHOP ON GEOGRAPHIC INFORMATION RETRIEVAL, 6., 2010, Zurich, Switzerland. *Proceedings...* New York, NY, USA: ACM, 2010. p. 1–8. Citado na página 31.

ALVARENGA, L.; DIAS, C. da C. Análise de domínio e gestão arquivística. *DataGramaZero*, [S.l], v. 13, n. 1, p. 7, 2012. Citado nas páginas 20, 21, 22, 24 e 29.

ALWIS, S. M. G. de; HIGGINS, S. E. Information as a tool for management decision making: a case study of Singapore. *Information Research*, Lund, Sweden, v. 7, n. 1, p. 114, 2001. Citado na página 29.

AMITAY, E. et al. Web-a-where: geotagging web content. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 27., 2004, Sheffield, United Kingdom. *Proceedings...* New York, NY, USA: ACM, 2004. p. 273–280. Citado nas páginas 30, 31 e 46.

ANASTÁCIO, I. *Location-Based Targeting and Ranking for Online Advertising*. 2009. 75 p. Dissertação (Master in Information Systems and Computer Engineering) — Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal, 2009. Citado nas páginas 18 e 46.

ANDRADE, L.; SILVA, M. J. Relevance ranking for geographic ir. In: WORKSHOP ON GEOGRAPHICAL INFORMATION RETRIEVAL, 3., 2006, Seattle, WA, USA. *Proceedings...* New York, NY, USA: ACM, 2006. Citado nas páginas 44 e 45.

AUER, S.; LEHMANN, J.; HELLMANN, S. Linkedgeodata: Adding a spatial dimension to the web of data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 8., 2009, Chantilly, VA, USA. *Proceedings...* Heidelberg, Germany: Springer, 2009. p. 731–746. Citado na página 44.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2. ed. Harlow, England: Addison-Wesley, 2011. Citado na página 43.

BAILEY, P. et al. The CSIRO enterprise search test collection. *ACM SIGIR Forum*, New York, NY, USA, v. 41, n. 2, p. 42–45, 2007a. Citado nas páginas 46, 51 e 64.

BAILEY, P. et al. Overview of the TREC 2007 enterprise track. In: TEXT RETRIEVAL CONFERENCE, 16., 2007, Gaithersburg, MD, USA. *Proceedings...* [S.l.]: NIST, 2007b. Citado na página 46.

BALOG, K. et al. Overview of the TREC 2008 enterprise track. In: TEXT RETRIEVAL CONFERENCE, 17., 2008, Gaithersburg, MA, USA. *Proceedings...* [S.l.]: NIST, 2008. Citado na página 46.

BORGES, K. A. V. et al. Discovering geographic locations in web pages using urban addresses. In: WORKSHOP ON GEOGRAPHICAL INFORMATION RETRIEVAL, 4., 2007, Lisboa, Portugal. *Proceedings...* New York, NY, USA, 2007. p. 31–36. Citado nas páginas 30 e 43.

BRÄSCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo, SP, Brasil. *Anais...* São Paulo: ANCIB, 2008. p. 1–14. Citado nas páginas 22 e 23.

BROUGHTON, V. The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings*, London, UK, v. 58, n. 1/2, p. 49–72, 2006. Citado nas páginas 22, 31, 32, 33 e 41.

BUCHER, B. et al. Geographic IR systems: requirements and evaluation. In: INTERNATIONAL CARTOGRAPHIC CONFERENCE, 22., 2005, A Coruña, Espanha. *Proceedings...* [S.l.], 2005. Citado nas páginas 45 e 46.

BUCKLEY, C.; VOORHEES, E. M. Retrieval evaluation with incomplete information. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 27., 2004, Sheffield, United Kingdom. *Proceedings...* New York, NY, USA: ACM, 2004. p. 25–32. Citado na página 45.

BUKOWSKA, E. et al. Ontology-based retrieval of experts: The issue of efficiency and scalability within the extraspec system. In: QUIRCHMAYR, G. et al. (Ed.). *Multidisciplinary Research and Practice for Information Systems*. Heidelberg, Germany: Springer, 2012. p. 272–286. Citado na página 29.

BUSCALDI, D.; ROSSO, P. A comparison of methods for the automatic identification of locations in wikipedia. In: WORKSHOP ON GEOGRAPHICAL INFORMATION RETRIEVAL, 4., 2007, Lisboa, Portugal. *Proceedings...* New York, NY, USA: ACM, 2007. p. 89–92. Citado na página 30.

CAMPOS, M. L. d. A. Lenguaje documentario: una perspectiva teórica. *Infodiversidad*, Buenos Aires, Argentina, v. 7, p. 95–114, 2004. Citado nas páginas 31, 32 e 33.

CARDOSO, N.; SANTOS, D. To separate or not to separate? reflections about current gir practice. In: WORKSHOP ON NOVEL METHODOLOGIES FOR EVALUATION IN INFORMATION RETRIEVAL, 1., 2008, Glasgow, UK. *Proceedings...* Heidelberg, Germany: Springer, 2008. Citado na página 30.

CARDOSO, N.; SOUSA, P.; SILVA, M. J. The University of Lisbon at GeoCLEF 2008. In: WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, CLEF, 9., 2008, Aarhus, Danmark. *Proceedings...* Heidelberg, Germany: Springer, 2008. Citado na página 46.

CHAVES, M. S. *Uma Metodologia para Construção de Geo-Ontologias*. 2009. 212 p. Tese (Informática) — Faculdade de Ciências. Universidade de Lisboa, Lisboa, Portugal, 2009. Citado na página 31.

CHOO, C. W. et al. Information culture and information use: An exploratory study of three organizations. *Journal of the American Society for Information Science and Technology*, New York, NY, USA, v. 59, n. 5, p. 792–804, 2008. Citado na página 29.

CHOU, S.-W. Knowledge creation: absorptive capacity, organizational mechanisms, and knowledge storage/retrieval capabilities. *Journal of Information Science*, London, UK, v. 31, n. 6, p. 453–465, 2005. Citado na página 30.

CRASWELL, N.; VRIES, A. P. de; SOBOROFF, I. Overview of the TREC 2005 enterprise track. In: TEXT RETRIEVAL CONFERENCE, 14., 2005, Gaithersburg, MD, USA. *Proceedings...* [S.l.]: NIST, 2005. v. 5, p. 199. Citado na página 45.

DIAS, C. da C.; ALVARENGA, L. Análise do domínio organizacional na perspectiva arquiviística: um estudo baseado na metodologia proposta por Designing and Implementing Recordkeeping Systems, DIRKS. *Ciência da Informação*, Brasília, DF, Brasil, v. 40, n. 2, p. 180–191, 2011. Citado na página 20.

DOLBY, J. et al. Extracting enterprise vocabularies using linked open data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 8., 2009, Chantilly, VA, USA. *Proceedings...* Heidelberg, Germany: Springer, 2009. p. 779–794. Citado na página 28.

DOMÈNECH, D. F. *Geographical Information Resolution and its Application to the Question Answering Systems*. 2007. 147 p. Tese (Intelligència Artificial) — Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, 2007. Citado na página 46.

EHLEN, P.; ZAJAC, R.; RAO, K. B. Location and relevance. In: INTERNATIONAL WORKSHOP ON LOCATION AND THE WEB, 2., 2009, Boston, Massachusetts, USA. *Proceedings...* New York, NY, USA: ACM, 2009. p. 1–3. Citado na página 44.

FERNANDEZ, M. et al. Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, [S.l.], v. 9, n. 4, p. 434–452, 2011. Citado na página 28.

FUJITA, M. S. A leitura documentária na perspectiva de suas variáveis: leitor-texto-contexto. *DataGramaZero*, [S.l], v. 5, n. 4, p. 1, 2004. Citado na página 22.

GARCÍA-CUMBRERAS, M. Á. et al. Information retrieval with geographical references: Relevant documents filtering vs. query expansion. *Information Processing & Management*, [S.l], v. 45, n. 5, p. 605–614, 2009. Citado na página 45.

GARDIN, J.-C. Document analysis and linguistic theory. *Journal of Documentation*, [S.l], v. 29, n. 2, p. 137–168, 1973. Citado nas páginas 22, 27 e 28.

GARFIELD, E. A tribute to SR Ranganathan, the father of indian library science: Part 1 - life and works. *Current Contents*, [S.l], v. 7, p. 37–44, 1984. Citado nas páginas 31 e 32.

GIESS, M.; WILD, P.; MCMAHON, C. The generation of faceted classification schemes for use in the organisation of engineering design documents. *International journal of information management*, [S.l], v. 28, n. 5, p. 379–390, 2008. Nenhuma citação no texto.

GIRGENSOHN, A. et al. Docubrowse: faceted searching, browsing, and recommendations in an enterprise context. In: INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES, 15., 2010, Hong Kong, China. *Proceedings...* New York, NY, USA: ACM, 2010. p. 189–198. Citado na página 44.

GOPINATH, M. Ranganathan's theory of facet analysis and knowledge representation. *DESIDOC Journal of Library & Information Technology*, [S.l], v. 12, n. 5, p. 16–20, 1992. Citado na página 32.

GOUVÊA, C. *Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web*. 2009. 88 p. Dissertação (Mestrado em Ciência da Computação) — Universidade Católica de Pelotas, Pelotas, RS, Brasil, 2009. Citado na página 31.

GUY, I. et al. Best faces forward: a large-scale study of people search in the enterprise. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 30., 2012, Austin, Texas, USA. *Proceedings...* New York, NY, USA: ACM, 2012. p. 1775–1784. Citado na página 29.

HALEVY, A. Y. et al. Enterprise information integration: successes, challenges and controversies. In: SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 31., 2005, Baltimore, USA. *Proceedings...* New York, NY, USA: ACM, 2005. p. 778–787. Citado nas páginas 22, 23 e 64.

HASSAN, A.; JONES, R.; DIAZ, F. A case study of using geographic cues to predict query news intent. In: SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 17., 2009, Seattle, Washington, USA. *Proceedings...* New York, NY, USA: ACM, 2009. p. 33–41. Citado nas páginas 30 e 31.

HJØRLAND, B. The classification of psychology: A case study in the classification of a knowledge field. *Knowledge Organization*, [S.l], v. 24, n. 4, p. 162–201, 1998. Citado nas páginas 24 e 32.

HJØRLAND, B. Domain analysis in information science: eleven approaches–traditional as well as innovative. *Journal of documentation*, [S.l], v. 58, n. 4, p. 422–462, 2002. Citado nas páginas 20, 21, 25, 26, 27, 28, 29, 32 e 64.

HJØRLAND, B. Is classification necessary after google? *Journal of Documentation*, [S.l], v. 68, n. 3, p. 299–317, 2012. Citado na página 22.

HJØRLAND, B.; ALBRECHTSEN, H. Toward a new horizon in information science: Domain-analysis. *Journal of The American Society for Information Science*, [S.l.], v. 46, n. 6, p. 400–425, 1995. Citado nas páginas 20 e 23.

HONG, M. Potential usage of faceted classification in internet information retrieval. *Interdisciplinary information sciences*, [S.l], v. 12, n. 1, p. 43–51, 2006. Citado nas páginas 43, 44 e 45.

HU, B.; SVENSSON, G. A case study of linked enterprise data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 9., 2010, Shanghai, China. *Proceedings...* Heidelberg, Germany: Springer, 2010. p. 129–144. Citado na página 28.

HU, X.; BANDHAKAVI, S.; ZHAI, C. Error analysis of difficult TREC topics. In: SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMAION RETRIEVAL, 26., 2003, Toronto, Canada. *Proceedings...* New York, NY, USA: ACM, 2003. p. 407–408. Citado na página 25.

JONES, C. B. et al. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, Bristol, PA, USA, v. 22, n. 10, p. 1045–1065, 2008. Citado na página 30.

JONES, R.; HASSAN, A.; DIAZ, F. Geographic features in web search retrieval. In: INTERNATIONAL WORKSHOP ON GEOGRAPHIC INFORMATION RETRIEVAL, 2., 2008, Napa Valley, CA, USA. *Proceedings...* New York, NY, USA: ACM, 2008. p. 57–58. Nenhuma citação no texto.

JOYCE, P. An analysis of the generic structure of customer service email. *Kinki University English Journal*, Osaka, Japan, v. 7, p. 37–53, 2011. Citado na página 27.

LA BARRE, K. Facet analysis. *Annual Review of Information Science and Technology*, [S.l], v. 44, n. 1, p. 243–284, 2010. Citado nas páginas 29, 32, 41, 44, 45 e 71.

LADEIRA, A. P. *Processamento de linguagem natural: caracterizacao da produção científica dos pesquisadores brasileiros*. 2010. 262 p. Tese (Doutorado em Ciência da Informação) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010. Citado nas páginas 27 e 28.

LEVELING, J.; HARTRUMPF, S. *University of Hagen at GeoCLEF 2007: Exploring Location Indicators for Geographic Information Retrieval*. Heidelberg, Germany: Springer, 2008. Citado na página 30.

LI, L. T.; TORRES, R. da S. Revisitando os Desafios da Recuperação de Informação Geográfica na Web. *Cadernos CPqD Tecnologia*, Campinas, SP, Brasil, v. 6, n. 1, p. 7–20, 2009. Citado na página 30.

LI, Z. et al. Indexing implicit locations for geographical information retrieval. In: WORKSHOP ON GEOGRAPHICAL INFORMATION RETRIEVAL, 3., 2006, Seattle, WA, USA. *Proceedings...* New York, NY, USA: ACM, 2006. Citado na página 30.

LIMA, G. . A. B. O Modelo Simplificado para Análise Facetada de Spiteri a partir de Ranganathan e do Classification Research Group (CRG). *Información, Cultura y Sociedad*, [S.l], v. 11, p. 57–72, 2004a. Citado nas páginas 32, 33, 34, 35, 36 e 38.

LIMA, G. A. B. de O. *Mapa Hipertextual (MHTX): um modelo para organização hipertextual de documentos*. 2004b. 207 p. Tese (Doutorado em Ciência da Informação) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2004b. Citado na página 33.

LIU, X. et al. Entity centric query expansion for enterprise search. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 21., 2012, Maui, Hawaii, USA. *Proceedings...* New York, NY, USA: ACM, 2012. p. 1955–1959. Citado na página 44.

LIU, X. et al. Finding relevant information of certain types from enterprise data. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 20., 2011, Glasgow, Scotland, UK. *Proceedings...* New York, NY, USA: ACM, 2011. p. 47–56. Citado na página 25.

LYKKE-NIELSEN, M. Domain analysis, an important part of thesaurus construction. *Advances in Classification Research Online*, [S.l], v. 11, n. 1, p. 9–50, 2011. Citado nas páginas 21 e 47.

MACULAN, B. C. M. d. S.; LIMA, G. A. B. d. O. Taxonomia facetada navegacional: agregando valor às informações disponibilizadas em bibliotecas digitais de teses e dissertações. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília, DF, Brasil. *Anais...* João Pessoa, PB, Brasil: ANCIB, 2011. Citado na página 44.

MARCELLA, R.; ILLINGWORTH, L. The impact of information behaviour on small business failure. *Information Research*, Lund, Sweden, v. 17, n. 3, p. 525, 2012. Citado na página 29.

NAVIGLI, R. Word sense disambiguation: a survey. *ACM Computing Surveys*, [S.l], v. 41, n. 2, p. 1–69, 2009. Citado na página 28.

NUNES, M. B. et al. Knowledge management issues in knowledge-intensive SMEs. *Journal of Documentation*, [S.l], v. 62, n. 1, p. 101–119, 2006. Citado na página 29.

O'FARRILL, R. T. Information literacy and knowledge management at work: Conceptions of effective information use at NHS24. *Journal of Documentation*, [S.l], v. 66, n. 5, p. 706–733, 2010. Citado na página 29.

OMELAYENKO, B. Integrating vocabularies: Discovering and representing vocabulary maps. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 1., 2002, Sardenha, Itália. *Proceedings...* Heidelberg, Germany: Springer, 2002. p. 206–220. Citado na página 28.

OREN, E.; DELBRU, R.; DECKER, S. Extending faceted navigation for rdf data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 5., 2006, Athens, GA, USA. *Proceedings...* Heidelberg, Germany: Springer, 2006. p. 559–572. Citado na página 44.

OVERELL, S. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. 2009. 181 p. Tese (Doutorado em Computação) — Imperial College London, London, UK, 2009. Citado na página 31.

OVERELL, S. E.; RÜGER, S. M. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, [S.l], v. 22, n. 3, p. 265–287, 2008. Citado na página 31.

PAGE, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web*. [S.l.], 1999. Disponível em: <http://ilpubs.stanford.edu:8090/422/>. Acesso em: 12 set. 2012. Citado na página 29.

PETERS, C. et al. Evaluating systems for multilingual and multimodal information access. In: WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, CLEF, 9., 2008, Aarhus, Danmark. *Proceedings...* Heidelberg, Germany: Springer, 2008. Citado na página 45.

PONTES, F. V. *Organização do conhecimento em bibliotecas digitais de teses e dissertações: uma abordagem baseada na classificação facetada e taxonomias dinâmicas*. 2013. 234 p. Tese (Doutorado em Ciência da Informação) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013. Citado na página 46.

PONTES, F. V.; LIMA, G. Â. B. d. O. Knowledge organization in digital environments: faceted classification theory applied. *Perspectivas em Ciência da Informação*, Belo Horizonte, MG, Brasil, v. 17, n. 4, p. 18–40, 2012. Citado na página 44.

RANGANATHAN, S. R. *Prolegomena to Library Classification*. New York: Asia Publishing House, 1967. Citado nas páginas 33, 58 e 66.

RULA, A. et al. On the diversity and availability of temporal information in linked open data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 11., 2012, Boston, USA. *Proceedings...* Heidelberg, Germany: Springer, 2012. p. 492–507. Citado na página 27.

SACCO, G. M. Research results in dynamic taxonomy and faceted search systems. In: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, 18., 2007, Regensburg, Germany. *Proceedings...* Washington, DC, USA: IEEE, 2007. Citado na página 44.

SAKAI, T.; KANDO, N. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, Heidelberg, Germany, v. 11, n. 5, p. 447–470, 2008. Citado nas páginas 45 e 46.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, [S.l], v. 24, n. 5, p. 513–523, 1988. Citado nas páginas 25 e 28.

SANTOS, D.; CARDOSO, N. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Lisboa, Portugal: Linguateca, 2007. Citado na página 46.

SANTOS, D. et al. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In: WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, CLEF, 9., 2008, Aarhus, Danmark. *Proceedings...* Heidelberg, Germany: Springer, 2008. p. 894–905. Citado na página 31.

SOLOMON, P. Bringing people, technology, and systems together through classification research: Designing, for change, learning, and maintenance. *Advances in Classification Research Online*, [S.l], v. 13, n. 1, p. 23–28, 2002. Citado nas páginas 23, 24, 26 e 29.

SOSNOVSKY, S.; DICHEVA, D. Ontological technologies for user modelling. *International Journal of Metadata, Semantics and Ontologies*, [S.l], v. 5, n. 1, p. 32–71, 2010. Citado na página 28.

SOUTHON, F. G.; TODD, R. J.; SENEQUE, M. Knowledge management in three organizations: An exploratory study. *Journal of the American society for Information Science and Technology*, New York, NY, USA, v. 53, n. 12, p. 1047–1059, 2002. Citado na página 26.

SPITERI, L. A simplified model for facet analysis: Ranganathan 101. *Canadian journal of information and library science*, [S.l.], v. 23, n. 1-2, p. 1–30, 1998. Citado nas páginas 32, 33, 35, 36, 37, 38, 39, 40 e 58.

TÁLAMO, M. Terminologia e documentação. *Tradterm Revista do Centro Interdepartamental de Tradução e Terminologia*, São Paulo, SP, Brasil, v. 7, p. 141–152, 2001. Citado nas páginas 22, 25, 26 e 29.

TODA, H. et al. Incorporating place name extents into geo-ir ranking. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 17., 2008, Napa Valley, California, USA. *Proceedings...* New York, NY, USA: ACM, 2008. p. 1489–1490. Citado na página 44.

VAKKARI, P.; JÄRVELIN, K. Explanation in information seeking and retrieval. In: SPINK, A.; COLE, C. (Ed.). *New directions in cognitive information retrieval*. Heidelberg, Germany: Springer, 2005. p. 113–138. Citado na página 33.

VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação ea difusão do conhecimento. *Ciência da Informação*, [S.l.], v. 31, n. 2, p. 152–162, 2002. Citado na página 29.

VICKERY, B. Faceted classification for the web. *Axiomathes*, Heidelberg, Germany, v. 18, n. 2, p. 145–160, 2008. Nenhuma citação no texto.

VOORHEES, E. The philosophy of information retrieval evaluation. In: WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, 2., 2001, Darmstadt, Germany. *Proceedings...* Heidelberg, Germany: Springer, 2002. p. 355–370. Citado na página 45.

WANG, Z.; CHAUDHRY, A. S.; KHOO, C. S. Using classification schemes and thesauri to build an organizational taxonomy for organizing content and aiding navigation. *Journal of documentation*, [S.l.], v. 64, n. 6, p. 842–876, 2008. Nenhuma citação no texto.

WILD, P. J.; GIESS, M. D.; MCMAHON, C. A. Describing engineering documents with faceted approaches: observations and reflections. *Journal of Documentation*, [S.l.], v. 65, n. 3, p. 420–445, 2009. Citado na página 71.

WU, H. C. et al. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, New York, NY, USA, v. 26, n. 3, p. 1–37, 2008. Citado na página 25.

YU, B.; CAI, G. A query-aware document ranking method for geographic information retrieval. In: WORKSHOP ON GEOGRAPHICAL INFORMATION RETRIEVAL, 4., 2007, Lisboa, Portugal. *Proceedings...* New York, NY, USA: ACM, 2007. p. 49–54. Citado nas páginas 45 e 46.