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

.

.

.

Resumo

Palavras-chave:

Abstract

Keywords:

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. (MACK, 2011)

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. (TANENBAUM; ZUCCHI, 2009)

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

Como demonstrado em (COSTA, 2001), muitos programas acabam por ter instruções redundantes ao longo de seu fluxo de execução. Assim, tempo computacional é perdido para se obter resultados já calculados.

Uma das técnicas propostas para reduzir esse desperdício de poder de processamento é a *DTM: Dynamic Trace Memoization*, ou Memorização Dinâmica de Traces. A DTM armazena o resultado de conjuntos de instruções executados anteriormente e, caso detecte uma execução redundante do mesmo conjunto, é capaz de armazenar os resultados e desviar o fluxo de controle para a instrução a ser executada após esse conjunto, substituindo a execução linear de cada instrução pelo resultado final, como se o bloco inteiro fosse uma instrução somente. A técnica será abordada com mais detalhes na seção **??**.

Em simulações realizadas por (COSTA, 2001), essa técnica foi capaz de aumentar o desempenho de programas do *SpecInt95 Benchmark Suite* de 1% até 21%, variando de acordo com o programa e os parâmetros utilizados na construção das unidades responsáveis por implementar o mecanismo DTM.

É possível então notar que há aplicações para as quais a implementação de uma unidade de DTM poderia melhorar significativamente o desempenho. Sendo assim, é interessante conhecer as os impactos desta para que seja possível melhor avaliar em que situações a utilização da DTM é proveitosa, considerando os *trade-offs* causados por sua presença.

1.2 Problema

A arquitetura de um circuito e a tecnologia utilizada para a sua geração estão intimamente ligados às características do circuito resultante. Considerando isso, algumas métricas utilizadas nesses circuitos resultantes servem para comparar apenas um dos fatores, seja uma mesma arquitetura em diversas tecnologias ou diversas arquiteturas em uma mesma tecnologia.

Segundo (CHU, 2006), as principais métricas que podem ser utilizadas nessas medições são área de chip, velocidade, consumo de potência e custo de produção. Esses quesitos são correlacionados, fazendo que alterações mudem os valores de mais de um ponto, senão todos.

O problema que motiva este trabalho é saber, considerando essas métricas, se a DTM é uma técnica viável para aplicações práticas. Caso seja, quais os *trade-offs* que o projetista deve considerar ao aplicar a DTM no seu projeto.

1.3 Objetivos

O objetivo deste trabalho é avaliar como as métricas citadas na seção 1.2 são alteradas com a implementação do mecanismo de DTM em um processador.

Mais especificamente, os objetivos podem ser descritos nos seguintes tópicos:

1. Implementar a memorização dinâmica de traços em uma arquitetura de processadores;
2. Produzir um circuito físico do processador e comparar os resultados da arquitetura padrão e da arquitetura com DTM nas seguintes métricas:

- área de chip;
- potência consumida;
- latência de ciclo;
- Executar programas de *benchmark* sobre as duas arquiteturas e comparar os resultados de performance de ambas.

1.4 Estrutura da tese

Esta tese está organizada em **??** capítulos. Esses capítulos foram ordenados em uma sequência lógica que facilitasse a compreensão do leitor, já que cronologicamente houve certo paralelismo durante a produção de algumas etapas.

- No capítulo 2 é demonstrado o processo metodológico adotado no desenvolvimento deste trabalho. Nele é descrito o processo de seleção da literatura base, a definição das ferramentas e parâmetros para a implementação da DTM e como foi planejado a análise e comparação dos resultados obtidos.
- Sequencialmente, no capítulo 3 é apresentada a fundamentação que serve como base teórica para o trabalho, incluindo uma conceituação mais apurada sobre a técnica de memorização dinâmica de traces, além de discorrer sobre a arquitetura de processadores escolhida como referência para a implementação e as tecnologias utilizadas para o desenvolvimento.

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;

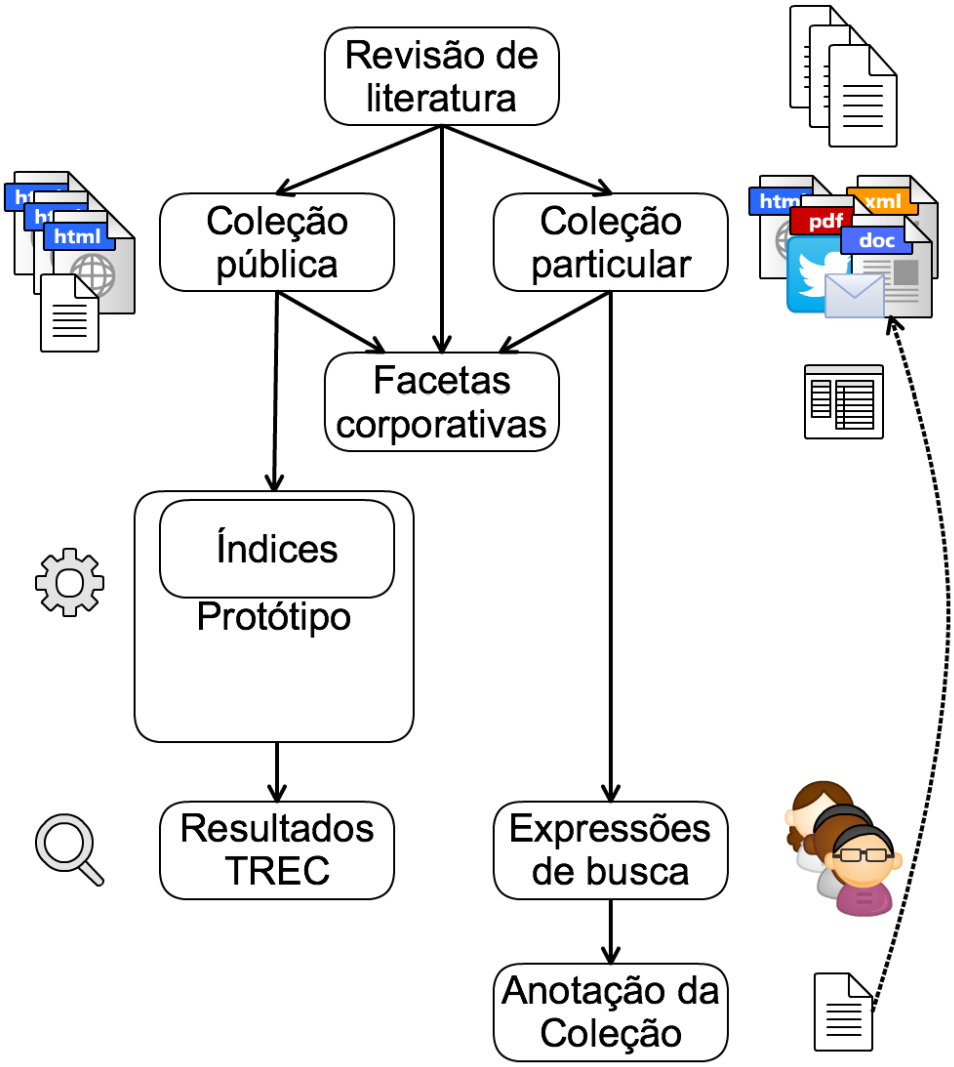
4. projetar e implementar um arcabouço de *software* que indexe a massa de dados de documentos corporativos usando estruturas de dados específicas – parte da pesquisa que pode ser classificada como estudo de caso;

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 **??**.

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 **??**.

A metodologia de prototipação se baseia no trabalho de **??**), 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 **??**, 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 (**??**), 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 (**??**).

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" (**??**, 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 (**??**). **??**) 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. (**??**)

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” (**??**). Também, **??**) 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 (**??**). Para a análise de domínio, esses são apenas subprodutos (**??**). 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” (**??**). 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 (**??**). 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 (**??**). 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 (**??**). 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 (**??**), linguagem documentária (**??**) ou simplesmente linguagem de indexação (**??**).

As linguagens de indexação representam a informação por alguma estrutura que reflete os objetivos do sistema de organização de conhecimento no qual é aplicada. **??**) 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 (**????**). Ao tentar localizar e compreender informação de múltiplos repositórios em organizações, por exemplo, **??**) 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 (**??**). Essa incompati-

bilidade 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 (**????**).

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 (**??**). 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 (**??**). 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 (**??**). Para **??**), 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 (**??**).

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 **??**), 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.

??) 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 (**??**).

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” (**??**). 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 classifi-

cação 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 (**??**). Segundo **??**), 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 **??**), 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 coocorrência de termos em documentos e na coleção (**??**). 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 (**????**).

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 (**??**). 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 (**??**). **??**, 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 (**??**, 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 (**??**, p. 25).

Para **??**, 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) (**??**).

Adicionalmente, há uma fração de evidências temporais que estão associadas à contemporaneidade implícita de dois eventos, indivíduos ou mesmo documentos. O reconheci-

mento 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 (**??**).

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 (**??**, 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" (**??**, p. 48), que é independente do idioma e portanto ultrapassa a informação provida pela sintaxe (**??**).

Exatamente por essa necessidade de conhecimento prévio (ou informação prévia), **??**, 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 psicolinguística, 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 (**??**), 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" (**??**, 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 (**????**). 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 (**??**). Isso explica-se por estruturas paradigmáticas serem compostas por relações lógicas não presentes no texto, ditada por convenções sociais ou profissionais (**??**, 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 (**??**), ontologias (**??????**) ou mesmo sistemas de informação estruturados. Em todos os casos, essas estruturas semânticas tentam servir como fonte de reconhecimento e desambiguação de significado (**??**), ou para extração (**??**) ou compatibilização de vocabulário corporativo (**??**).

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 (**??**). 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 (**??**). No entanto, essas estratégias dependem da disponibilidade e riqueza dos metadados usados e de um controle de vocabulário muito maior do que tem ocorrido (**????**).

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 (**????**).

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 (**??**) e webométricos (**??**). 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 (**??**). 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 (**????**). 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 (**??????????**). 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 (**??**).

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 (**??**), tais como nomes de lugar, nomes de empresas, códigos de endereçamento postal ou endereços (**????**), sejam esses indicadores explícitos ou implícitos (**??**).

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 (**??**). 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 **??**), 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 (**????**).

Trabalhos têm usado repositórios que contenham dados semiestruturados ou estruturados, como *gazetteers* ou a Wikipedia (**??**), ou reconhecem entidades geográficas em um conjunto restrito de documentos, como notícias de jornais, onde há vocabulário e estrutura gramatical controlados (**??**). 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 (**??**). 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 **??????**) e **??**) 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 (**????**), algo incompatível com o ambiente corporativo. Por fim, o emprego de

ontologias (**??**), 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 (**??**), 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 (**??**). 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 (**??**). 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 (**????**).

Feita a análise de um domínio, a partir de uma amostra representativa de seus documentos, 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 (**??**, 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 **????**) e **??**) 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 (**??????**).

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 (**??**, 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 (**??**, p. 62).

As facetas mais gerais do domínio são também conhecidas como categorias fundamentais, ou apenas categorias (**??**). 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 (**??**). Essas categorias fundamentais poderiam ser então subdivididas em subfacetas ou mesmo expandidas para novas facetas em função da necessidade do classificador (**??**). 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), *By-product* (subproduto), *Space* (espaço) e *Time* (tempo) (**??**). 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 (**??**).

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 (**??**). 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 (**??**). 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) (**??**, 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 (**??**).

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 (**??**). 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 (**??**, p. 61).

A teoria da análise facetada de **??**) 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 (**??**). 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" (**??**, p. 64). Essa condição motivou o trabalho de **??**) que pode ser sintetizado, em suas próprias palavras, como "um modelo de análise facetada consolidado e simples de seguir" (**??**, Sec. 2, tradução nossa).

As principais diferenças entre os trabalhos de Ranganathan e do CRG podem ser consultadas em fontes variadas como **??????**) e **??**), cabendo a esta seção apresentar apenas a visão simplificada por **??**) 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

??, 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. **??**) 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” (**??**, 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 **??**) 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 (**??**, 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

??) 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 (**??**, 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 (**??**, 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. **??**) e **??**) 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” (**??**, 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 (**??**, 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” (**??**, Sec. 3, tradução nossa). Foi proposto como um cânone por Ranganathan e foi mantido pelo CRG. Porém, **??**, 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 **??**) e **??**) 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” (**??**, Sec. 3, tradução nossa). Ou, “um cachorro de raça ‘Dálmata’ será sempre um dálmata” (**??**, p. 67). A explicação parece tão incompatível com a definição dada para este princípio que **??**, 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 habi-

tat? 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*" (**??**, 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 (**??**). Foi proposto apenas pelo CRG, diferindo do postulado das cinco categorias fundamentais de Ranganathan. Segundo **??**), 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" (**??**, 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. **??**) 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 classificação. 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 **??**). 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 **??**), porém merece ser explicado como parte do princípio da terminologia usual, embora esteja em conflito com este princípio, como definido por **??**). Pelo cânone da reticência a terminologia usada não pode refletir um viés da visão de um grupo de classificadores ou qualquer preconceito. **??**) 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 (**??**). Complementar ao anterior, o *princípio de homônimo*, também proposto por Ranganathan, assegura que um número de classe da notação representa apenas um único assunto (**??**).

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 **??**), 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 (**??**).

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.

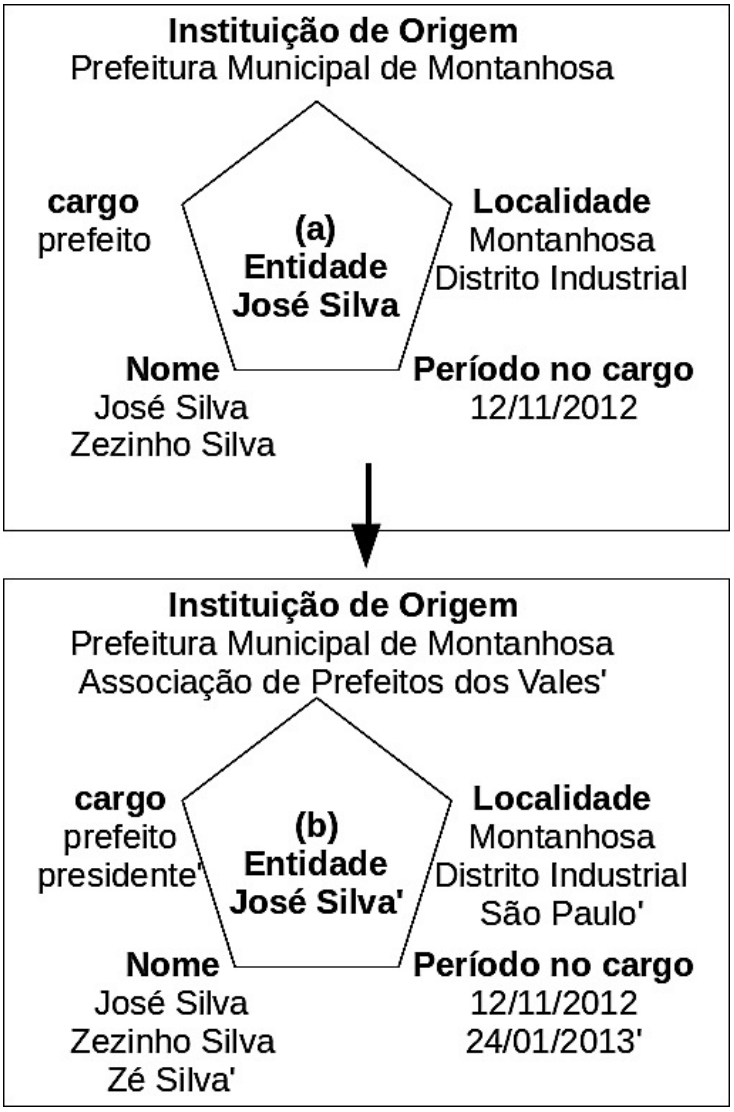
3.2.3 Uso de facetas na busca e na recuperação de informação

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 (**??**). 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 (**??**). 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

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

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 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 (**??????**). 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 (**??**).

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 (**??**). 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 (**??**, 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 (**??**). 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 (**??????????**). Porém, "sem um compromisso com vocabulários controlados, resta pouca esperança de que as tecnologias da Web

semântica alcancem seu fim" (**??**, 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.

??, 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 (**??**, 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 (**????**), especialmente no momento da consulta (**??**). 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, **??**) e **??**) 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 (**??**) e responda com eficiência ao usuário de informação (**????**).

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 (**????**). 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 (**??????????**), algumas delas bem aceitas na indústria e na academia. Porém, há limitações para avaliação de sistemas corporativos pelas mesmas metodologias, dadas suas especificidades (**??**).

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 (**??**). 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 (**??**).

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 (**??????**), visando reconhecer evidências espaciais e linguísticas. Metodologia similar, porém específica para sistemas em língua portuguesa, é denominada Avaliação de Reconhecimento de Entidades Mencionadas (HAREM) (**??**).

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 (**??**), 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 (**??????**). 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 (**????**).

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 **??**) e **??**) 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 **??**). 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 (**??**).

Referências

CHU, P. P. *RTL hardware design using VHDL: coding for efficiency, portability, and scalability*. [S.l.]: John Wiley & Sons, 2006. Citado na página 6.

COSTA, A. T. da. *Explorando dinamicamente o reuso de traces em nível de arquitetura de processador*. 2001. Tese (Doutorado) — COPPE/UFRJ, Rio de Janeiro, 2001. Citado nas páginas 5 e 6.

MACK, C. A. Fifty years of moore's law. *IEEE Transactions on semiconductor manufacturing*, IEEE, v. 24, n. 2, p. 202–207, 2011. Citado na página 5.

TANENBAUM, A. S.; ZUCCHI, W. L. *Organização estruturada de computadores*. [S.l.]: Pearson Prentice Hall, 2009. Citado na página 5.