UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

WALBER ANTONIO RAMOS BELTRAME

UM SISTEMA DE DISSEMINAÇÃO SELETIVA DA INFORMAÇÃO BASEADO EM CROSS-DOCUMENT STRUCTURE THEORY

VITÓRIA
2011

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

WALBER ANTONIO RAMOS BELTRAME

UM SISTEMA DE DISSEMINAÇÃO SELETIVA DA INFORMAÇÃO BASEADO EM CROSS-DOCUMENT STRUCTURE THEORY

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Informática da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Mestre em Informática, na área de concentração em Informática na Educação, sob orientação do professor Doutor Davidson Cury e co-orientação do professor Doutor Crediné Silva de Menezes.

VITÓRIA
2011

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Beltrame, Walber Antonio Ramos, 1983-
B453s Um sistema de disseminação seletiva da informação baseado em Cross-Document Structure Theory / Walber Antonio Ramos Beltrame. – 2011.
87 f. : il.

Orientador: Davidson Cury.
Corientador: Crediné Silva de Menezes.
Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Disseminação seletiva da informação. 2. Recuperação da informação. 3. Processamento de linguagem natural (Computação). 4. Teoria dos grafos. 5. Sistemas de recuperação da informação. I. Cury, Davidson. II.Menezes, Crediné Silva de, 1952-. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 004

Um Sistema de Disseminação Seletiva da Informação baseado em Crossdocument Structure Theory

Walber Antonio Ramos Beltrame

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

Aprovada em 30/08/2011 por:

Prof. Dr. Davidson Cury - DI/UFES

Prof. Dr. Crediné Silva de Menezes - DI/UFES

Prof. Dr. Orivaldo de Lira Tavares - DI/UFES

Prof. Dr. Alberto Nogueira de Castro Junior - UFAM

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Vitória-ES, agosto de 2011

DEDICATÓRIA

A minha esposa e a minha filha.

AGRADECIMENTOS

Agradeço ao grande mestre Dede por tudo, a Juliana Kowata, pela ajuda e discussões sobre a dissertação. Aos amigos do laboratório, aos meus familiares, em especial, aos meus pais. Agradeço também a todos os professores do curso, com os quais muito aprendi.

RESUMO

Um Sistema de Disseminação Seletiva da Informação é um tipo de Sistema de Informação que visa canalizar novas produções intelectuais, provenientes de quaisquer fontes, para ambientes onde a probabilidade de interesse seja alta. O desafio computacional inerente é estabelecer um modelo que mapeie as necessidades específicas de informação, para um grande público, de modo personalizado. Para tanto, é necessário mediar à estruturação da unidade informacional, de maneira que contemple a pluralidade de atributos a serem considerados pelo processo de seleção de conteúdo.

Em recentes publicações acadêmicas, são propostos sistemas baseados em marcação de dados sobre textos (modelos de meta-dados), de forma que o tratamento da informação manifesta-se entre computação de dados semi-estruturados e mecanismos de inferência sobre meta-modelos. Tais abordagens utilizam-se apenas da associação da estrutura de dados com o perfil de interesse. Para aperfeiçoar tal característica, este trabalho propõe a construção de um sistema de disseminação seletiva da informação baseado em análise de múltiplos discursos por meio da geração automática de grafos conceituais a partir de textos, concernindo à solução também os dados não estruturados (textos).

A proposta é motivada pelo modelo *Cross-Document Structure Theory*, recentemente difundido na área de Processamento de Língua Natural, voltado para geração automática de resumos. O modelo visa estabelecer correlações de natureza semântica entre discursos, por exemplo, se existem informações idênticas, adicionais ou contraditórias entre múltiplos textos. Desse modo, um dos aspectos discutidos nesta dissertação é que essas correlações podem ser usadas no processo de seleção de conteúdo, o que já fora evidenciado em outros trabalhos correlatos. Adicionalmente, o algoritmo do modelo original é revisado, a fim de torná-lo de fácil aplicabilidade.

ABSTRACT

A System for Selective Dissemination of Information is a type of information system that aims to harness new intellectual products, from any source, for environments where the probability of interest is high. The inherent challenge is to establish a computational model that maps specific information needs, to a large audience, in a personalized way. Therefore, it is necessary to mediate informational structure of unit, so that includes a plurality of attributes to be considered by process of content selection.

In recent publications, systems are proposed based on text markup data (meta-data models), so that treatment of manifest information between computing semi-structured data and inference mechanisms on meta-models. Such approaches only use the data structure associated with the profile of interest. To improve this characteristic, this paper proposes construction of a system for selective dissemination of information based on analysis of multiple discourses through automatic generation of conceptual graphs from texts, introduced in solution also unstructured data (text).

The proposed model is motivated by Cross-Document Structure Theory, introduced in area of Natural Language Processing, focusing on automatic generation of summaries. The model aims to establish correlations between semantic of discourse, for example, if there are identical information, additional or contradictory between multiple texts. Thus, an aspects discussed in this dissertation is that these correlations can be used in process of content selection, which had already been shown in other related work. Additionally, the algorithm of the original model is revised in order to make it easy to apply.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Um Sistema de Disseminação Seletiva da Informação é um tipo de Sistema de Informação que visa canalizar novas produções intelectuais, provenientes de quaisquer fontes, para ambientes onde há alta probabilidade de interesse. Esse conceito origina-se da proposição de (LUHN, 1961), que sistematiza serviços de notificação de acordo com perfis.

O serviço estabelecido tornou-se comum em bibliotecas digitais, voltados à produção de listas selecionadas de títulos e à distribuição de resumos das novas aquisições. Com a evolução das tecnologias de rede de computadores e das formas de comunicação, o recurso consolidou-se como padrão de sistemática capaz de divulgar atualizações entre diferentes plataformas e sítios de conteúdo (ALMEIDA, 2008).

A estruturação do interesse é um dos focos principais desse campo de pesquisa, aliado aos mecanismos de correlação entre os critérios estabelecidos e significância dos documentos. Para este trabalho, o termo “interesse” refere-se à necessidade de um agente humano ou de sistema de computador em adquirir informação que corresponda a alguma condição particular, relevante e atual: novos interesses podem surgir ao longo do tempo, assim como necessidades antigas se tornam inválidas. Segundo (SOUTO, 2008) (SOUTO, 2008), os sistemas devem prover pró-atividade quanto à identificação dos diferentes contextos dessa dinamicidade.

O desafio computacional inerente é estabelecer um modelo que mapeie as necessidades específicas de informação, para um grande público, de modo constante e personalizado. Para tanto, é necessário mediar à estruturação da unidade informacional (SOUTO, 2008) (SOUTO, 2008), de maneira que contemple a pluralidade de atributos a serem considerados pelo processo de seleção de conteúdo.

Refere-se a atributo de seleção de conteúdo como àquele que dita uma propriedade, pelo qual estabelece sentido ao objeto: o título de um livro, a data de um evento ou a quantidade de rotações do motor de um automóvel, enfim, toda qualidade passível de interesse.

De acordo com (SOUTO, 2008) (SOUTO, 2008), além do interesse, os sistemas de disseminação seletiva devem atentar-se para critérios igualmente importantes: qualidade, veracidade, síntese, interface com especialistas humanos e interoperabilidade.

Adiciona-se a essas, por iniciativa deste texto, a capacidade de redução do esforço cognitivo, a facilidade de interação e o estímulo a construção de conhecimento, tratados no decorrer da dissertação.

1.1 Hipótese

Em recentes publicações acadêmicas – cita-se (ALMEIDA, 2008), (EIRÃO, 2009), (MORALES-DEL-CASTILLO, PEDRAZA-JIMÉNEZ, *et al.*, 2009), (KANSA e BISSELL, 2010) e (EIRÃO, 2011) – são propostos sistemas de disseminação seletiva da informação baseados em marcação de dados sobre textos (modelos de meta-dados), de forma que o tratamento da informação manifesta-se entre computação de dados semi-estruturados e mecanismos de inferência sobre meta-modelos (ver Seções 2.1).

Tais abordagens utilizam-se apenas da associação da estrutura de dados com o perfil de interesse (ver Seção 2.5) e, em muitos casos, de forma a não contemplar todas possíveis combinações (SOUTO, 2008). Isso vai ao encontro de (GUIZZARDI, 2005), ao relatar falhas de interoperabilidade, de suporte metodológico e de expressividade da linguagem utilizada na marcação de dados.

Essa hipótese conduziu esta pesquisa a outras indagações:

i. Os atuais sistemas de disseminação seletiva baseados em marcação de meta-dados, que são comumente utilizados na comunidade científica (EIRÃO, 2011), podem ser aperfeiçoados quanto ao perfil de interesse, se apoiados por outras tecnologias. Este trabalho propõe a utilização de análise de múltiplos discursos, por meio da geração automática de grafos conceituais a partir de textos (KOWATA, 2010), concernindo à solução também os dados não estruturados (textos) (ver Seção 4.3);

ii. O modelo *Cross-Document Structure Theory* (RADEV, 2000), recentemente difundido na área de Processamento de Língua Natural e voltado para geração

automática de resumos (sumarização) – (HASSAN, RADEV, *et al.*, 2009), (QAZVINIAN e RADEV, 2010) e (JBARA e RADEV, 2011) – propõe a identificação de correlações de natureza semântica entre discursos: em múltiplos textos podem existir informações idênticas, adicionais ou contraditórias, dentre outras. Essas correlações podem ser usadas no processo de seleção de conteúdo, o que já fora evidenciado em outros trabalhos correlatos (JORGE, 2010), em que operadores de seleção são utilizados para compor um único texto com frases provenientes de múltiplas fontes (ver Seção 3.4). Ainda que recentes, as propostas se mostram interessantes para experimentação nos sistemas de disseminação seletiva, o que é base deste trabalho.

iii. De acordo com o que é exposto em (RADEV, 2000), (ZHANG e RADEV, 2004) e (KRISHNA, HASSAN, *et al.*, 2011), o modelo lida com um conjunto reduzido de textos, após um processo de organização de documentos (*clustering*), e identifica as relações semânticas por meio de bases anotadas manualmente. A fim de torná-lo de fácil aplicabilidade, o algoritmo do modelo original deve ser revisado.

Parte-se de um campo de pesquisa ainda não explorado em sistemas de disseminação: as metodologias segundo a perspectiva pragmático-discursiva têm evoluído para contribuir no processo de elucidação do contexto social dos indivíduos.

O diálogo constitui uma interação verbal de alto valor, que se bem explorado, é capaz de estabelecer, além das relações intrínsecas aos textos, uma visão de mundo do interlocutor e de como esse olhar é aceito pelo ambiente. Desse modo, a função dos sistemas de disseminação deve ir além do papel de propagar informação, mas de ser um mecanismo promotor da interação continuada, voltado à construção de novos conteúdos, logo, de novos conhecimentos.

1.2 Objetivos

O objetivo geral deste trabalho é evoluir recentes propostas de sistemas de disseminação seletiva da informação para cenários em que a seleção de conteúdo seja vista como suporte para construção contínua de conhecimento, à medida que transfere o foco centrado na sistematização de dados para uma análise baseada em correlações entre discursos.

O objetivo específico do trabalho é propor um modelo conceitual e arquitetural de sistema de disseminação que se apoie nessa perspectiva. O mérito maior esperado, além da constatação da validade das hipóteses, é corroborar com futuros trabalhos ao evidenciar as dificuldades e os objetos utilizados para superá-las.

1.3 Metodologia

Este trabalho utilizou-se, como metodologia científica, das etapas de revisão de literatura dos principais conceitos inerentes e da concepção de soluções para os objetos enunciados na hipótese, caracterizando-a como exploratória e descritiva.

1.4 Estrutura da dissertação

Os próximos capítulos estão estruturados da seguinte forma:

No Capítulo 2 é feita uma revisão sobre sistemas de disseminação seletiva da informação, sendo apresentados os principais conceitos tratados no trabalho. É feita uma avaliação subjetiva de quatro sistemas relevantes.

No Capítulo 3 é referenciada a teoria que titula o trabalho, a *Cross-Document Structure Theory*, de forma evidenciar as argumentações que serão utilizadas nos capítulos seguintes.

No Capítulo 4 é proposto um modelo Vetorial estendido para identificação de correlações semânticas entre discursos. Tal proposta é motivada por dificuldades de aplicação dos algoritmos atualmente consolidados.

No Capítulo 5 é relatado o resultado dos estudos de casos sobre o modelo vetorial, assim como são expostos os artefatos gerados na concepção de um sistema de disseminação seletiva da informação, denominado de Fique Sabendo.

No Capítulo 6 são apresentadas as verificações dos objetivos alcançados, as considerações finais e os possíveis trabalhos futuros.

2 SISTEMAS DE DISSEMINAÇÃO SELETIVA DA INFORMAÇÃO

Segundo (SOUTO, 2008) (SOUTO, 2008), o conceito “disseminação seletiva da informação” pode ser definido como um serviço que se utiliza de perfis (individuais ou de grupo) explícitos ou implícitos (ver Seção 2.2) para submeter periodicamente (ou disponibilizar acesso a) um pacote de informações resultantes de seleção, realizada por ação humana ou por tecnologia.

Os sistemas computacionais de disseminação seletiva da informação automatizam esses serviços. O funcionamento padrão dos sistemas de disseminação pode ser descrito como um conjunto de atividades sequenciais e cíclicas (LUHN, 1961):

i. Percorrer as fontes produtoras da informação (cadastradas de alguma forma) ou varrer a base de dados que contenham as novas informações submetidas;

ii. Normalizar e indexar as novas formas informações, por meio de descritores, no repositório de dados. A escolha dos descritores (ver Seção 2.1) está diretamente relacionada com a estruturação do perfil de interesse, ora estático ora evolutivo;

iii. Estabelecer ou recuperar os perfis de interesse (ver Seção 2.2);

iv. Selecionar por meio de pesquisa, de casamento de padrão ou de critérios pré-definidos (ver Seção 2.3) os documentos relevantes aos perfis recuperados;

v. Apresentar (disseminar) os resultados da seleção em formatos entendíveis pelos requerentes;

vi. Permitir avaliação dos resultados pelos requerentes e retroalimentar o sistema para melhoria contínua da composição dos perfis.

Nas próximas seções são relatadas as três principais diretrizes que estabelecem um sistema de disseminação seletiva: a forma com que um conteúdo deve ser estruturado, ainda que proveniente de diversas origens; o registro e a manipulação do interesse (necessidade) personalizável e a etapa de seleção de documentos.

Posteriormente são discutidos sistemas relevantes da literatura atual, de modo que são exemplificados os três conceitos anteriores. Por fim, são mostrados os desafios e as questões norteadoras da temática.

2.1 Formas de representação do conteúdo

Para (LANCASTER, 2004), a definição de formatos para estruturar conteúdos baseia-se na escolha de atributos (ou descritores) específicos que tornem possível a correspondência com os perfis de interesse. Desse modo, o processo de representação passa-se por duas etapas: obtenção dos descritores e fomento desses índices. As formas de obtenção dos descritores são enumeradas a seguir:

i. Os descritores são fornecidos pelos produtores do conteúdo, previamente acordados com os sistemas de disseminação e formalizados em modelo estrutural;

ii. Os índices são extraídos do dado original, desde que os sistemas de disseminação conheçam o formato do dado de origem e possuam algoritmos de conversão;

iii. Os atributos são gerados automaticamente por meio de inferência, utilizando-se de algoritmos de classificação ou de organização de dados (*clustering*).

O fomento, ou a etapa de estruturação dos dados, tem por objetivo facilitar o posterior casamento dos padrões, entre índices e perfis. O formato final dos atributos dependerá da estratégia de comparação de dados adotada. Destacam-se alguns padrões:

i. Para dados estruturados: organização em sistemas de bancos de dados (PARSAYE, CHIGNELL, *et al.*, 1989);

ii. Para dados textuais: representação na forma de vetores de frequência dos termos (palavras) do documento (SALTON, WONG e YANG, 1975);

iii. Para dados hiper-textuais: utilização dos modelos associativos em rede (AGOSTI e MARCHETTI, 1992);

iv. Para meta-dados: integração entre modelos conceituais e linguagens de descrição (GUIZZARDI, 2005).

Uma política de qual padrão adotar passa-se pela avaliação dos requisitos arquiteturais em que o sistema de disseminação será concernido. Diante da possibilidade de comunicação com outros sistemas, deve-se optar por aquele que proverá maior capacidade de se interoperar dados. Logo, o padrão de meta-dados tem maior aceitabilidade.

2.2 Formas de representação do interesse

A representação do interesse (necessidade da informação) é o fator de maior relevância em sistemas de disseminação, uma vez que é o principal critério para se determinar o que deve ser selecionado e, posteriormente, disseminado. Segundo (SOUTO, 2008) (SOUTO, 2008), o interesse se apresenta sobre as óticas:

i. Interesse externo: os sistemas restringem as opções de interesse para os elementos contidos nos formatos de representação de conteúdo, de forma organizada por temas e por classes. É a forma mais simples de representação, em que se guardam somente as informações de quais categorias serão escolhidas;

ii. Interesse explícito: os atributos (conjunto de dados que representa um interesse) são informados pelo receptor, que é consciente das necessidades que possui e as tornam explícitas, facilitando o trabalho dos sistemas de disseminação quanto aos critérios que devem ser avaliados;

iii. Interesse implícito: os atributos podem ser inferidos por meio da percepção de que o receptor possui uma necessidade, mas não a manifesta. Sistemas que lidam com interesse implícito devem possuir modelos de dados do receptor (conhecidos como modelos do usuário), em que são manipuladas as informações necessárias para a inferência, por exemplo, o histórico das solicitações, opções feitas no sistema e os próprios dados do receptor (nome, endereço, sexo, idade, profissão, qualificações, relacionamento com outros receptores, etc.).

Quando os interesses não possuem correspondentes nos formatos de representação, os sistemas de disseminação podem ignorar tal fato e não apresentar nenhum suporte ao receptor. Ou então, adotar estratégias de aprendizado e evolução do modelo de conteúdo, buscando correlacionar os atributos, até que torne a necessidade mapeada.

Usualmente, os interesses são expressos no padrão de linguagem de consulta, de modo a ser interpretada pelos sistemas de seleção, semelhante às pesquisas personalizadas em sistemas de recuperação da informação (BAEZA-YATES e RIBEIRO-NETO, 1999).

Dessa forma, estratégias de expansão de consulta também são utilizadas para melhorar a qualidade dos resultados (SPARCK-JONES, 1992). Tais técnicas se utilizam de dicionário de palavras, tesauro e semi-ontologia para aumentar a quantidade de termos da pesquisa.

2.3 Formas de seleção da informação

A seleção da informação é realizada basicamente pela comparação entre a representação do conteúdo e a representação do interesse. Essa comparação pode ser exata, selecionando somente os documentos que satisfaçam o interesse, ou parcial, selecionando também conteúdos similares ao interesse. Os sistemas parciais trabalham de acordo com o princípio da incerteza das necessidades, decorrente da própria subjetividade na formulação do que é de interesse ou não.

Outra forma de seleção é aquela que observa as relações entre os próprios documentos, principalmente nos de conteúdo associativo (hiper-texto). Nesse formato de seleção, outros documentos, além daqueles que satisfazem a consulta, são selecionados por possuírem alguma correlação notória.

Em (MONTEIRO, 2009), é relatado outro meio de seleção de documentos, centrado no perfil do receptor e nos demais perfis, ao que se designou seleção social, em que a relevância de um recurso condiz com quantos outros interesses similares existem. Em sistemas de disseminação que permitem esse tipo de interação, é comum a seleção baseada em indicações de conteúdo, num processo de construção coletiva de perfis comunitários.

Após o processo de seleção do conteúdo, a tarefa dos sistemas de disseminação é prover o acesso a informação selecionada. Nesse ponto, as questões que merecem destaque são quanto às formas de comunicação e quanto às formas apresentação dos resultados:

i. Em relação à comunicação, os sistemas podem adotar estratégias assíncronas, disponibilizando a informação em momento mais adequado, visto que o interesse do receptor pode não ser iminente a requisição. Outra questão relacionada à comunicação é definir qual protocolo será utilizado, dentre as várias possibilidades, por exemplo, serviços de mensagem eletrônica (*e-mail*), comunicação móvel e etc.;

ii. Em relação à apresentação dos resultados, os sistemas podem agregar à solução mecanismos de pré-visualização dos documentos selecionados, a fim de evitar que os receptores acessem informações indesejadas.

2.4 Avaliação de sistemas relevantes

A intenção deste tópico é analisar alguns sistemas de disseminação, publicados na última década. Foram selecionados quatro sistemas relevantes, descritos nos parágrafos a seguir, em ordem crescente cronológica. Os critérios de avaliação estão descritos na Tabela 1. Para os últimos critérios da tabela, foram utilizadas como referencial teórico: Qualidade, Veracidade, Síntese (SARACEVIC, 1996); Interface de mediação, Interoperabilidade (SOUTO, 2008) (SOUTO, 2008); Redução do esforço cognitivo (LAZARTE, 2000); Facilidade de interação (PRIMO, 2007); Construção do Conhecimento (FREIRE, 1999).

Tabela 1. Critérios de avaliação utilizados para avaliação dos sistemas

Quanto à forma de representação do conteúdo		
Forma de obtenção	(1) Fornecimento	Quando os descritores são fornecidos
	(2) Extração	Os índices são extraídos do conteúdo
	(3) Geração	Os descritores são inferidos
Fomento	(4) Estrutural	Utiliza-se de bancos de dados
	(5) Vetorial	Utiliza-se de modelos vetoriais
	(6) Rede	Utiliza-se de modelos de rede
	(7) Semântico	Utiliza-se de modelos de meta-dados
Quanto à forma de representação do interesse		
Tipo de Interesse	(8) Externo	Interesse dirigido pelo conteúdo
	(9) Explícito	Interesse é expresso pelo receptor
	(10) Implícito	Interesse é inferido pelo sistema

Abordagem	(11) Simples	Não utiliza processos evolutivos
	(12) Complexa	Utiliza evolução ou expansão do interesse
Quanto à seleção da informação		
Tipo de seleção	(13) Exata	Os resultados correspondem à pesquisa
	(14) Parcial	Os resultados são similares à pesquisa
	(15) Relacional	Seleção devido à relação entre documentos
	(16) Social	Seleção devido a critérios sociais do receptor
Outros critérios		
(17) Qualidade		O sistema avalia resultados (retroalimentação)
(18) Veracidade		O sistema valida índices e documentos
(19) Síntese		O sistema sintetiza os resultados
(20) Interface de mediação		O sistema possui interface para mediação humana
(21) Interoperabilidade		O sistema integra facilmente com outros sistemas
(22) Redução do esforço cognitivo		A curva de aprendizagem do sistema é baixa
(23) Facilidade de interação		A interação é mediada e intuitiva
(24) Construção do conhecimento		O sistema auxilia na aprendizagem construtiva

A verificação dos critérios, neste trabalho, é subjetiva e analítica, realizada por meio de leitura e de interpretação dos textos referenciais publicados. A intenção não é classificar os sistemas, mas expor de maneira organizada como os conceitos são abordados. Os sistemas de disseminação seletiva avaliados por este trabalho estão descritos a seguir de forma resumida e o resultado da avaliação sintetizado na Tabela 2.

i. MySDI (FERREIRA e SILVA, 2001): arquitetura genérica para projetar serviços de disseminação. O modelo estrutura-se em quatro camadas – do usuário, da informação, da classificação e de filtragem. São elaborados agentes de software (não é evidenciado qual padrão de modelo de agentes utilizado) que se interagem, para estabelecer coordenação entre níveis. Na camada de classificação são utilizadas máquinas de inferência, para geração de índices. Para manipulação destes, são utilizadas máquinas vetoriais. O perfil de interesse é constituído por interações (navegação) em sítios de conteúdo marcados por temas: cada ação do usuário alimenta o sistema de forma positiva ou negativa quanto à temática de interesse. A seleção de conteúdo é feita por mecanismos de consulta vetorial, mas existe um módulo de verificação de correspondência entre perfis, que é utilizado como elemento de cálculo de similaridade. As formas de comunicação definidas pelo sistema são as usuais providas pelos protocolos *Web*. O sistema não é dirigido a conhecimento, restringindo-se a ser uma ferramenta de disseminação;

ii. SemCast (PAPAEMMANOUIL e ÇETINTEMEL, 2004): proposta de um sistema baseado em difusão altamente distribuída, para fluxos de grande volume de dados. Propõe uma abordagem semântica para filtragem de conteúdo. Canais menores de interesse são gerados dinamicamente e correlacionados na forma de topologia de rede, sendo que os descritores são previamente estabelecidos. O interesse é do tipo externo, dirigido pelo mapeamento semântico. Não é relatado formalmente no texto como mudanças de interesse são tratadas, mas se conclui que a abordagem é do tipo simples, visto que o tipo de interesse é externo (não evolutivo). O tipo de seleção é a exata – o interesse é expresso pela assinatura de canais. O sistema não é dirigido a conhecimento, restringindo-se a ser uma ferramenta de disseminação.

iii. SABiO (BAX, ALVARENGA, *et al.*, 2004): sistema de disseminação voltado a bibliotecas, organizado em três agentes de software (não é evidenciado qual padrão de modelo de agentes utilizado): agente de captura, agente de interface e agente de notificação. Os descritores são fornecidos, basicamente expressos como dados específicos sobre livros e publicações. O fomento dos dados é estrutural, em banco de dados relacional. O tipo de interesse é explícito, em que o usuário informa parâmetros de consultas (do tipo booleano). A abordagem é simples, visto que nenhum mecanismo evolutivo é proposto. O tipo de seleção é a exata, feito com rotinas para consulta a banco de dados relacional. Não é dito se existem formas de comunicação para interoperabilidade. O sistema não é dirigido a conhecimento, restringindo-se a ser uma ferramenta de disseminação.

iv. G-ToPSS (PETROVIC, LIU e JACOBSEN, 2005): proposta de um sistema baseado em difusão altamente distribuída, voltado a escalabilidade de padrões anotados semanticamente. O padrão anunciado é o da *Web Syndication* (tecnologia que define formatos de marcação XML[1] / RSS, acrônimo para *RDF Site Summary*, *Really Simple Syndication* ou *Rich Site Summary*). Os índices são fornecidos pelo mapeamento semântico. O tipo de interesse é externo e a abordagem é simples. O tipo de seleção é exato. O sistema não é dirigido a conhecimento, restringindo-se a ser uma ferramenta de disseminação.

[1] www.w3.org/XML/

Após essa análise, verificou que os critérios estabelecidos não são exclusivos, ou seja, pode haver sistemas híbridos, que contemplam dois ou mais conceitos de mesma categoria, por exemplo, se possuírem como formas de obtenção de índice tanto por fornecimento quanto por extração.

Desse modo, para pontuar cada trabalho, esta dissertação definiu pesos escalares para avaliação de um conceito: inicialmente possui peso zero; se existir referências sobre o assunto no texto, soma-se "um"; se o texto aprofundar o tema e apresentar novas propostas, também se soma. Ao final, a avaliação do conceito poderá ter um valor entre zero e dois. A ordenação dos conceitos segue a Tabela 1, que foi estruturada para revelar o grau de atendimento de critérios adotados para se avaliar as referências sobre sistemas de disseminação.

Tabela 2. Síntese dos resultados de avaliação dos sistemas de disseminação

	Critério	**i**	**ii**	**iii**	**iv**	**∑**
1	Quando os descritores são fornecidos	0	2	1	2	**5**
2	Os índices são extraídos do conteúdo	2	0	0	0	**2**
3	Os descritores são inferidos	1	0	0	0	**1**
4	Utiliza-se de bancos de dados	0	0	2	0	**2**
5	Utiliza-se de modelos vetoriais	2	0	0	0	**2**
6	Utiliza-se de modelos de rede	1	2	0	1	**4**
7	Utiliza-se de modelos de meta-dados	0	2	0	2	**4**
8	Interesse dirigido pelo conteúdo	1	2	0	2	**5**
9	Interesse é expresso pelo receptor	1	1	2	1	**5**
10	Interesse é inferido pelo sistema	2	0	0	0	**2**
11	Não utiliza processos evolutivos	0	2	2	2	**6**
12	Utiliza evolução ou expansão do interesse	1	0	0	0	**1**
13	Os resultados correspondem à pesquisa	0	2	2	2	**6**
14	Os resultados são similares à pesquisa	2	0	0	0	**2**
15	Seleção devido à relação entre documentos	1	2	0	2	**4**
16	Seleção devido a critérios sociais do receptor	1	0	1	0	**2**
17	O sistema avalia resultados (retroalimentação)	1	0	0	0	**1**
18	O sistema valida índices e documentos	1	0	0	0	**1**
19	O sistema sintetiza os resultados	1	2	0	2	**5**
20	O sistema possui interface para mediação humana	1	0	2	0	**3**
21	O sistema integra facilmente com outros sistemas	1	2	0	2	**5**
22	A curva de aprendizagem do sistema é baixa	1	1	1	1	**4**
23	A interação é mediada e intuitiva	1	0	2	1	**4**
24	O sistema auxilia na aprendizagem construtiva	0	0	0	0	**0**
	Total	**22**	**20**	**15**	**20**	

2.5 Evidências e desafios a superar

O preposto de que os sistemas atuais de disseminação não são rigorosos para questões de perfis de interesse e para a construção do conhecimento é confirmado pelos resultados, conforme o que se observa na avaliação dos critérios dez, onze, doze e o último da tabela anterior. Ainda que seja pequena a porção de sistemas avaliados, pode-se assumir que há uma generalização dessa corrente (EIRÃO, 2011).

Tal característica pode ter origem no fato de que os últimos esforços científicos para esse campo de sistemas focam-se na melhoria de desempenho e no avanço tecnológico, principalmente nos sistemas que se utilizam de meta-dados (OLIVEIRA, 2009).

Diante o exposto, existem desafios a superar quanto aos sistemas de disseminação em relação às abordagens centradas nos interesses e nas necessidades, não somente de dados e de informações, mas em novas possibilidades de socialização.

Uma evolução dos sistemas e do aporte tecnológico não invalidará a magnitude de conceitos como interoperabilidade e facilidade de uso, que continuarão coexistindo. No entanto, ressalta-se que mais estudos sob a ótica da construção do conhecimento são necessários.

2.6 Considerações parciais

Neste capítulo foram abordados os principais conceitos relevantes para esta dissertação sobre os sistemas de disseminação seletiva da informação. Quatro sistemas atuais foram analisados, segundo os critérios revistos por este trabalho.

A partir dessa avaliação, foram expostas as evidências e os desafios que serviram como motivação para continuidade dos próximos capítulos. No capítulo seguinte é realizada a fundamentação teórica sobre o modelo que serviu de base para a solução apresentada neste trabalho.

3 CROSS-DOCUMENT STRUCTURE THEORY

O objetivo deste capítulo é discutir o modelo linguístico computacional *Cross-Document Structure Theory* – CST (RADEV, 2000) – que visa estabelecer relações de natureza semântico-discursiva (identidade, similaridade, contradição, temporalidade) entre unidades informativas textuais de diferentes documentos. O resultado do algoritmo é usado em operadores de seleção para compor um único documento resumido (sumarização automática). Originalmente, o trabalho é inspirado nas seguintes referências:

i. (TRIGG, 1983) (TRIGG e WEISER, 1986): propõe um modelo de relacionamento entre sentenças textuais, baseado na composição de tipos básicos de ligações, de maneira que estabelece se uma sentença é uma argumentação, uma contradição ou um cenário de outra sentença. Os tipos são definidos com base em propriedades identificadas por padrões linguísticos mapeados;

ii. (MANN e THOMPSON, 1987): define o modelo linguístico discursivo RST (*Rhetorical Structure Theory*), sendo uma metodologia para análise do discurso que propõe o agrupamento de frases satélites em torno de uma frase central, ou núcleo. O núcleo relaciona-se com seus satélites e com outros núcleos por meio de relações definidas pelo modelo.

O modelo CST propõe uma metodologia para representação de relações entre unidades textuais, definindo tais relações. Para o modelo, um documento é composto de parágrafos, um parágrafo é composto de sentenças, uma sentença é composta de sintagmas e um sintagma é composto de palavras. Uma unidade textual será quaisquer dessas: ou o documento todo, ou um parágrafo e assim sucessivamente.

As correlações se estabelecem entre qualquer nível de unidade, formando um grafo de relações, em que cada nó representa a unidade informativa textual e as arestas representam as relações entre elas. No modelo original (RADEV, 2000) foram propostas 24 relações. Para a Língua Portuguesa do Brasil, o conjunto de relações foi refinado para 14 relações e classificados em categorias de relações (JORGE, 2010). Para melhor entendimento, as relações foram ilustradas neste trabalho na Tabela 3.

Tabela 3. Relações CST, sentenças e exemplos

Relações	Sentenças	Exemplos
Identity	(1) é idêntico a (2)	1 - Avião cai e mata 17 pessoas. 2 - Avião cai e mata 17 pessoas.
Equivalence	(1) é equivalente a (2)	1 - Avião cai e mata 17 pessoas. 2 - Avião cai e 17 pessoas morrem.
Translation	(1) é tradução de (2)	1 - Hello World! 2 - Olá Mundo!
Subsumption	(1) contêm complemento de (2)	1 - Com três anos, sou uma criança. 2 - Eu tenho três anos.
Contradiction	(1) é uma contradição de (2)	1 - O trânsito estava calmo. 2 - O Trânsito estava intenso.
Background	(1) contêm um histórico de (2)	1 - Ela já divorciou pela 4ª vez. 2 - Ontem ela se divorciou.
Modality	(1) é uma indicação de que (2)	1 - Sou o mais rico do bairro. 2 - Comprei a única mansão daqui.
Attribution	(1) contêm a fonte de (2)	1 - Paulo fez gol, diz jornal. 2 - Paulo fez gol.
Summary	(1) é um resumo de (2)	1 - Chutei a bola. 2 - Eu chutei a redonda bola.
Follow-up	(1) é um fato posterior a (2)	1 - A energia foi restabelecida. 2 - Faltou energia na cidade.
Elaboration	(1) contêm detalhe de algo em (2)	1 - 1,5% são analfabetos. 2 - A minoria não saber ler.
Indirect speech	(1) menciona algo de (2)	1 - Ele disse que iria ganhar. 2 - Ele disse: “irei ganhar!”.
Overlap	(1) contêm fato novo de (2)	1 - Ele entrou e Ela saiu. 2 - Ele entrou e a viu.
Citation	(1) contêm uma citação de (2)	1 - Ela disse o que ele falou: “irei”. 2 - Ele falou: “irei”.

A tipologia definida em (JORGE, 2010), reproduzida na Tabela 4, organiza as relações em duas categorias principais: de conteúdo, que agrupa relações primárias como similaridade, complementaridade e contradição; e de apresentação, que define aspectos secundários da informação, como a atribuição de autoria e identificação de traduções para outras línguas. Essa subdivisão é uma iniciativa de minimizar a ambiguidade e a subjetividade que as relações possam transmitir. Porém, algumas questões conceituais evidenciam deficiências dessas definições, como mostrado nos próximos parágrafos.

Tabela 4. Classificação das relações CST – adaptado de (JORGE, 2010)

<table>
<tr><td rowspan="8">Conteúdo</td><td rowspan="5">Redundância</td><td rowspan="3">Total</td><td>Identity</td></tr>
<tr><td>Equivalence</td></tr>
<tr><td>Summary</td></tr>
<tr><td rowspan="2">Parcial</td><td>Overlap</td></tr>
<tr><td>Subsumption</td></tr>
<tr><td rowspan="3">Complemento</td><td rowspan="2">Temporal</td><td>Background</td></tr>
<tr><td>Follow-up</td></tr>
<tr><td>Atemporal</td><td>Elaboration</td></tr>
<tr><td colspan="2">Contradição</td><td>Contradiction</td></tr>
<tr><td rowspan="4">Apresentação</td><td colspan="2" rowspan="3">Autoria</td><td>Modality</td></tr>
<tr><td>Citation</td></tr>
<tr><td>Attribution</td></tr>
<tr><td colspan="2" rowspan="2">Estilo</td><td>Translation</td></tr>
<tr><td>Indirect speech</td></tr>
</table>

(AFANTENOS, DOURA, *et al.*, 2004) relatam que o modelo carece de embasamento e os equívocos podem estar relacionados com a pragmática que envolve o discurso. Além disso, pode-se notar que certos tipos de relações possuem definições semelhantes, ocasionado ambiguidade:

i. No exemplo da Tabela 3, é dito que o trânsito estava calmo e depois é dito que o trânsito estava intenso. As sentenças são contraditórias, visto que os predicados são antagônicos. Mas, podem transmitir ideia de temporalidade, se ora o trânsito estava lento, outrora estava calmo (ainda que o tempo verbal das frases seja o mesmo);

ii. São expressas relações temporais de passado (*Background*) e de futuro (*Follow-up*), assumindo que algum texto retrata o presente. No entanto, se uma unidade textual aborda algo sobre o passado de outro texto, esse será o futuro da anterior, ou seja, sempre se constatam as duas relações (passado e futuro). O mais adequado seria adotar somente um tipo de relação que retrata quando duas unidades textuais ocorrem em tempos diferentes (ordem cronológica);

iii. Nas relações de redundância e de complemento, a relação que aborda resumo (*Summary*) possui as mesmas definições de complemento, porém na ordem inversa.

Enquanto complemento é algo adicional, o resumo é a abstração. Do raciocínio do item anterior, se um texto é resumo de outro, esse será o complemento;

Não obstante a isso, (RADEV, 2000) propõe que as relações sejam binárias (ou existe a relação ou não existe) e (JORGE, 2010) que sejam excludentes (não pode haver relações de mesmo nível de categoria entre duas unidades). Este trabalho propõe que as correlações sejam transcritas na forma de probabilidade ou de ponderação, ou seja, ao assumir que textos possuem uma relação, que essa seja mensurável. Por exemplo, na frase "O trânsito estava lento" e na frase "O trânsito estava calmo", é possível que haja 50% de contradição e 50% de temporalidade. Do ponto de vista computacional, essa abordagem é útil e adequada para lidar com a subjetividade inerente (ver Capítulo 4).

A formalização das relações, em (RADEV, 2000), é feita por meio de textos livres, não se fundamentando por modelos lógicos, uma vez que a concepção da metodologia é voltada para a construção de bases anotadas por agentes humanos, logo, a explicação textual facilitara o entendimento.

Por conseguinte, (RADEV, 2000) expôs que é possível a criação de métodos para obtenção das relações de forma automática, por meio de algoritmos de aprendizagem e técnicas de computação de linguagem natural, após anotação de coleções por especialistas humanos.

3.1 Identificação automática das relações

O método voltado à classificação automática de relações CST, detalhado em (ZHANG e RADEV, 2004), opera sobre um conjunto de processamentos sequenciais, tendo como pré-requisitos: construção de coleções manualmente anotadas (ZHANG, OTTERBACHER e RADEV, 2003) e treinamento dos classificadores não lineares (FREUND e SCHAPIRE, 1997). Realizada a aprendizagem de máquina, segue-se:

i. Para novos textos, ainda não rotulados, é feito um agrupamento (*clustering*) por meio de algoritmos estatísticos, a fim de refinar a coleção em pequenos conjuntos com alta probabilidade de existência das relações;

ii. Para cada conjunto, são realizados operações sobre os textos para determinação de estruturas lexicais, sintáticas e semânticas. As estruturas são parâmetros avaliados pelo classificador;

iii. A máquina realiza a classificação das sentenças, por correspondência de padrões, gerando o grafo CST daquele conjunto de textos.

Conforme (ZHANG e RADEV, 2004), os resultados encontrados não foram satisfatórios. Visando aperfeiçoar a eficácia, (MURAKAMI, NICHOLS, *et al.*, 2009), (KAWAHARA, INUI e KUROHASHI, 2010) e (MURAKAMI, NICHOLS, *et al.*, 2010) adicionaram ao processo outras técnicas de análise do discurso, como alinhamento estrutural (BROWN, LAI e MERCER, 1991). Cada relação foi tratada por classificadores diferenciados, utilizando-se de máquinas vetoriais (ver Capítulo 4). As soluções foram experimentadas em larga escala, obtendo melhores resultados.

Todavia, a necessidade de bases anotadas é um fator que reduz a aplicabilidade do método, visto que a tarefa de classificação manual exige esforços de especialistas. Aliado a isso, a subjetividade conceitual das relações podem ocasionar ruídos indesejáveis (AFANTENOS, DOURA, *et al.*, 2004), limitando a atuação dos classificadores.

3.2 Operadores de seleção de conteúdo

Além das relações CST, (RADEV, 2000) propôs etapas para sumarização: os textos são estruturados internamente após processo de análise de estruturas lexicais, sintáticas e semânticas; após essa etapa de análise, as relações CST são estabelecidas e as unidades textuais relacionadas são organizadas no grafo; na última etapa do método, o conteúdo é selecionado de acordo com a informação dada pelas relações.

O autor propõe também, na etapa de seleção, utilizar operadores de preferência (Figura 1). Para um operador de contradição, por exemplo, as sentenças relacionadas por meio da relação *Contradiction* terão uma preferência maior.

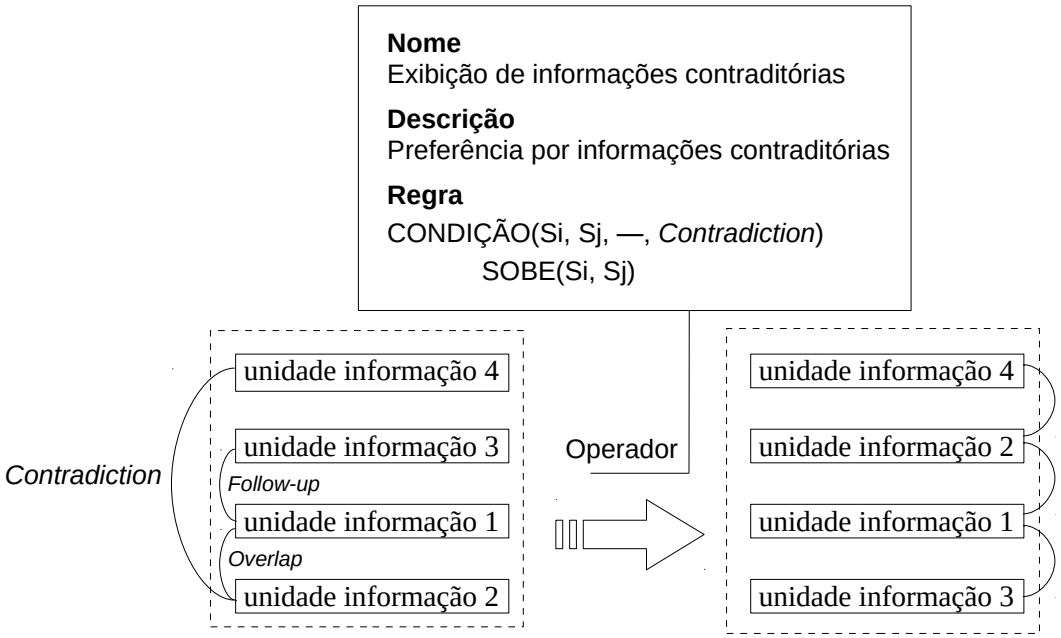

Figura 1. Exemplo de operador de seleção – adaptado de (JORGE, 2010)

Nessa última ilustração, Si e Sj são as unidades de informação, ou sentenças, que são parâmetros avaliados nas condições dos operadores, ou seja, se a relação existir, o método de subir posições é aplicado à lista.

Em (JORGE, 2010), são propostos: operador de apresentação de informação contextual; operador de apresentação de eventos que evoluem no tempo; operador de identificação de autoria; operador de redução de redundância e operador de exibição de informações contraditórias. É manifestada a possibilidade de execução de vários operadores, por ordem inversa de preferência e a criação de novos operadores.

3.3 Evidências e desafios a superar

A organização e o refinamento das relações CST estagiam entre a evolução teórica da metodologia e a validação experimental dos métodos de obtenção automática. As propostas que se utilizaram de máquinas vetoriais obtiveram melhores resultados, mas ainda relata a necessidade de *corpus*, o que envolve um esforço muito grande. Portanto, outro norte voltado à aplicabilidade é desejável. Para o uso em sistemas de disseminação seletiva, em que é comum grande volume de dados, as soluções vetoriais indicam-se como apropriadas.

3.4 Considerações parciais

Apesar da incipiência do modelo e da necessidade de revisões conceituais, este trabalho considera que a proposta de seleção de múltiplos textos por meio de correlação semântica entre discursos, em consonância com a CST (ver Seção 4.2), poderá ser aplicada para melhoria das formas de seleção de conteúdo em sistemas de disseminação seletiva, desde que novos métodos assistam aos requisitos desses sistemas (ver Tabela 1 e Capítulo 6).

4 FORMALIZAÇÃO DE UM MODELO VETORIAL ESTENDIDO

De acordo com (BAEZA-YATES e RIBEIRO-NETO, 1999), modelos de recuperação da informação podem ser formalizados como a quádrupla [D, Q, F, $R(q_i, d_j)$]:

i. D é um conjunto composto pelas visões lógicas dos documentos na coleção, chamadas tipicamente de representações;

ii. Q é um conjunto composto pelas necessidades de informação do usuário, chamadas tipicamente de consultas;

iii. F é um conjunto de ferramentas (*framework*) para lidar com as representações dos documentos, com as consultas e com os relacionamentos entre esses;

iv. $R(q_i, d_j)$ é uma função de ordenação que associa um número real à uma consulta q_i (pertencente a Q) e uma representação do documento d_j (pertencente a D).

Outra questão importante, tratada em (BAEZA-YATES e RIBEIRO-NETO, 1999), é a prática de se realizar operações sobre os textos dos documentos, dentre elas pode-se citar: eliminação de palavras indesejadas (*stopwords*); utilização de radicais léxicos dos termos (*stemming*); utilização de substantivos (eliminação de adjetivos, advérbios e verbos).

Para os sistemas de disseminação seletiva da informação, a formalização também é válida caso adaptações sejam feitas ao modelo: D é a representação dos novos documentos (ver Seção 2.1); Q é a representação do interesse (ver Seção 2.2); e F é a representação da seleção (ver Seção 2.3).

Dentre os modelos clássicos, tem-se o Vetorial (SALTON, WONG e YANG, 1975). Nesse modelo os documentos são representados como vetores no espaço *n*-dimensional, onde *n* é o total de termos índices (palavras) de todos os documentos no sistema. As consultas também são representadas como vetor de termos da pesquisa. Para calcular a similaridade entre eles é adotada alguma função matemática vetorial, geralmente, cosseno (Tabela 5):

$$cos(\vec{V}, \vec{E}) = \frac{\vec{V}\vec{E}}{|\vec{V}||\vec{E}|}$$

No modelo (Figura 2), o valor escalar de cada dimensão de um vetor é determinado por métodos de ponderação. A principal função desse método é o aumento da eficácia da recuperação, que depende de dois fatores: os documentos que poderão ser relevantes às necessidades do usuário devem ser recuperados e os itens que poderão ser irrelevantes devem ser rejeitados. Pode-se, por meio dessa diretriz, estabelecer as seguintes heurísticas:

i. *Term frequency* (tf): um documento que menciona um termo de consulta com mais frequência estará mais relacionado com a consulta e, portanto, deve receber uma pontuação mais elevada:

$tf_{i,j} = n_{i,j} / \sum_k n_{k,j}$

$n_{i,j} =$ frequência do termo *i* no documento *j*

ii. *Inverse document frequency* (idf): expressões que acontecem em quase todos os documentos não são úteis para diferenciá-los e, portanto, é necessário introduzir um mecanismo para atenuar os efeitos dos termos que muito ocorrem:

$idf_{i,j} = n_{i,j} \log(D / \sum_d t_i)$

$n_{i,j} =$ frequência do termo *i* no documento *j*

D = números de documentos na coleção

$\sum_d t_i$ = quantidade de documentos onde a frequência do termo *i* é maior que zero

iii. *Term discrimination* (tf.idf): sugere que as condições ideais são aquelas capazes de distinguir (discriminar) os documentos do restante da coleção e, portanto, obtido pelo produto da frequência do termo pelo inverso do documento:

$tf_{i,j} \times idf_{i,j}$

Tabela 5. Exemplificação do modelo Vetorial

Sejam os documentos D1, D2 e D3:
D1 - "O governador Mário Pereira solicitou bolsa agrícola para famílias"
D2 - "Ministro da Agricultura esteve com Governador do Paraná"
D3 - "O Governador do Paraná quer investir na agricultura"
Logo, $D = 3$.

Sejam as consultas Q1, Q2 e Q3:
Q1 - "governador Paraná agricultura"
Q2 - "Paraná quer melhor agricultura"
Q3 - "Mário Paraná quer investir na bolsa de valores"

Removendo as palavras indesejadas: "o", "a", "de", "do", "da", "para", "com", "na" e a forma minúscula das palavras, tem-se:

Termos	$n_{i,j}$			$1/\sum_k n_{k,j}$	$\log(D/\sum_d t_i)$	$tf_i \times idf_i$	$n_{i,j}$		
	D1	**D2**	**D3**				**Q1**	**Q2**	**Q3**
"governador"	1	1	1	1/3 = 0.3	log(3/3) = 0	0	1	0	0
"mário"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	1
"pereira"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"solicitou"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"bolsa"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	1
"agrícola"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"famílias"	1	0	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"ministro"	0	1	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"agricultura"	0	1	1	1/2 = 0.5	log(3/2) = 0.2	0.1	1	1	0
"esteve"	0	1	0	1/1 = 1	log(3/1) = 0.5	0.5	0	0	0
"paraná"	0	1	1	1/2 = 0.5	log(3/2) = 0.2	0.1	1	1	1
"quer"	0	0	1	1/1 = 1	log(3/1) = 0.5	0.5	0	1	1
"investir"	0	0	1	1/1 = 1	log(3/1) = 0.5	0.5	0	0	1
	$w_{i,j}$ = n(tf.idf)			$\lvert D_j\rvert = \sqrt{\sum_i w_{i,j}^2}$			$w_{i,j}$ = n(tf.idf)		
"governador"	0	0	0				0.5	0	0
"mário"	0.5	0	0	$\lvert D1\rvert = \sqrt{6 \times 0.5^2} = 1.2$			0	0	0.5
"pereira"	0.5	0	0	$\lvert D2\rvert = \lvert D3\rvert = 0.7$			0	0	0
"solicitou"	0.5	0	0	$\lvert Q1\rvert = \lvert Q2\rvert = 0.5 ; \lvert Q3\rvert = 1.0$			0	0	0
"bolsa"	0.5	0	0	$Q \cdot D_j = \sum w_{i,j} \times q_{i,j}$			0	0	0.5
"agrícola"	0.5	0	0	$cos(Q1, D1) = 0/0.6 = 0$			0	0	0
"famílias"	0.5	0	0	$cos(Q1, D2) = 0.2/0.4 = 0.5$			0	0	0
"ministro"	0	0.5	0	$cos(Q1, D3) = 0.2/0.4 = 0.5$			0	0	0
"agricultura"	0	0.1	0.1	$cos(Q2, D1) = 0/0.6 = 0$			0.1	0.1	0
"esteve"	0	0.5	0	$cos(Q2, D2) = 0.2/0.4 = 0.5$			0	0	0
"paraná"	0	0.1	0.1	$cos(Q2, D3) = 0.3/0.4 = 0.8$			0.1	0.1	0.1
"quer"	0	0	0.5	$cos(Q3, D1) = 0.5/1.2 = 0.4$			0	0.5	0.5
"investir"	0	0	0.5	$cos(Q3, D2) = 0/0.7 = 0$ $cos(Q3, D3) = 0.3/0.7 = 0.4$			0	0	0.5
Maior relevância: $Q1 \rightarrow (D2, D3)$; $Q2 \rightarrow D3$; $Q3 \rightarrow (D1, D3)$									

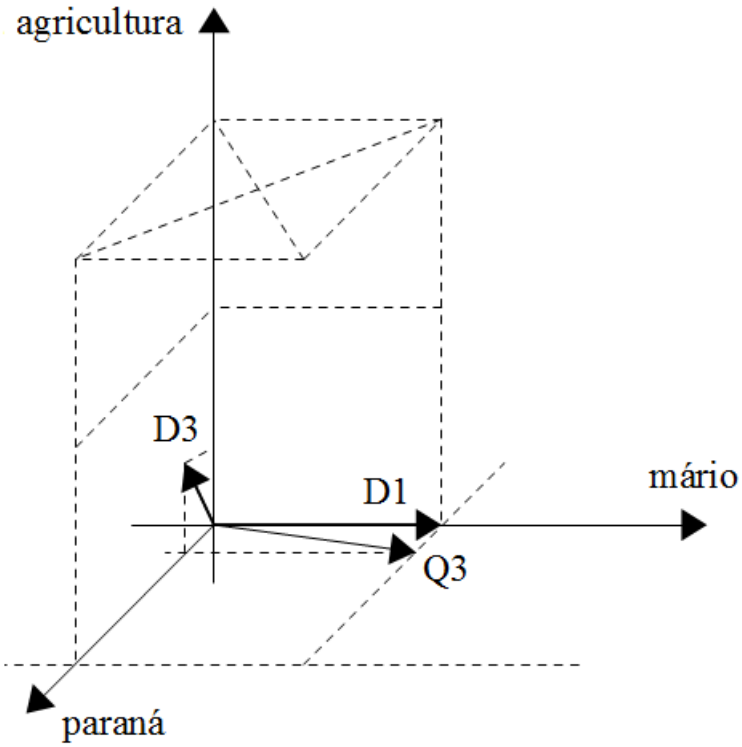

Figura 2. Ilustração do modelo Vetorial

No exemplo da consulta Q3 da Tabela 5, os resultados não foram precisos e nota-se uma limitação do modelo Vetorial em tratar polissemia, ou seja, termos que podem ser usados para expressar coisas diferentes em contextos diferentes (“bolsa” relacionada à “economia” – “bolsa de valores” – e “bolsa” relacionada ao “benefício” – “bolsa agrícola”). A escolha por indexação de termos é outro fator limitante, em que o uso por sintagmas poderia ser mais adequado, observado em “Mário Paraná”, “Governador do Paraná” e “Mário Pereira”. Por último, outra deficiência comum no modelo é a questão de palavras diferentes com significados próximos, visto em “agrícola” e “agricultura”.

4.1 Modelos vetoriais estendidos

O modelo Vetorial pode ser estendido ou adaptado, com base em outras três técnicas, para aumentar a semântica e o contexto na representação dos documentos, a fim de melhorar a eficiência (BAEZA-YATES e RIBEIRO-NETO, 1999). Citam-se:

i. Booleano Vetorial: adiciona-se uma linguagem de relacionamento booleano entre os termos da consulta: *e*, *ou*, *não*. As funções de similaridade contemplam tais condições (BEIGBEDER, 2005);

ii. Indexação semântica latente: questiona-se a significância da palavra-chave como candidata a índice. Estabelece casamento conceitual entre documentos e consultas, por meio da redução do espaço vetorial, em operações algébricas (DEEWESTER, DUMAIS, *et al.*, 1990);

iii. Baseado em tópicos: questiona-se a independência dos índices (vetores ortogonais), considerando que certos conceitos são relacionados. Compõe o espaço vetorial por meio de tesauros e de ontologias, selecionando as entidades representativas como dimensões (BECKER, 2003);

iv. Baseado em redes neurais (máquina vetorial): utiliza-se do casamento de padrões entre consultas e documentos, em que cada pesquisa resulta num sinal que ativa os termos índice, em interações sucessivas. O conjunto resposta é definido por meio desse processo e poderá conter documentos que não compartilham nenhum termo-índice da consulta, mas que tenham sido ativados durante o processo (HEARST, DUMAIS, *et al.*, 1998).

Por iniciativa desta dissertação, é investigada uma adaptação do modelo Vetorial que, baseado em recente trabalho sobre geração automática de grafos conceituais a partir de textos (KOWATA, 2010), possa suportar a detecção automática das relações CST. A abordagem é explicada nas próximas seções.

4.2 Um modelo vetorial estendido baseado em grafos conceituais

Segundo (DIESTEL, 2005), grafo é uma estrutura matemática, definida por $G = (V, E)$, sendo E (arestas) o conjunto que representa as associações entre os elementos do conjunto V (vértices). Em (SOUZA, BOERES, *et al.*, 2006), é estabelecido equivalência entre mapas conceituais e grafos.

(KOWATA, 2010) define uma abordagem híbrida sobre manipulação de textos, em Língua Portuguesa do Brasil, para produzir mapas conceituais. Apesar de o intuito ser de apoio ao processo de ensino e aprendizagem, é provido uma etapa intermediária de geração de grafos manipuláveis computacionalmente, explicada na seção seguinte.

4.2.1 Reconhecimento de grafos conceituais a partir de textos

De acordo com (KOWATA, 2010), o reconhecimento de mapas conceituais a partir de texto (expresso em Português do Brasil) é a capacidade de se representar um documento *d* por meio de um mapa conceitual *mc*.

Sendo o documento *d* constituído por sentenças $s_1...s_n$, ou seja, $d = \{s_1, s_2, ..., s_n\}$, para cada sentença s_i existem proposições $p_i...p_n$ a serem extraídas que possibilitam a construção de um mapa conceitual *mc* (KOWATA, 2010). Uma proposição p_i é definida por um conjunto de três elementos ordenados, c_{1i}, r_i e c_{2i}, nos quais c_{1i} e c_{2i} são conceitos e r_i uma relação entre esses conceitos.

Percebe-se que tais definições remontam a concepção clássica de “sujeito” e “predicado”, constituintes da “oração”. Em (KOWATA, 2010), entende-se “conceito” como um número reduzido de palavras que definem “uma regularidade percebida em objetos e eventos” e “relação” como uma proposição rotulada entre conceitos.

Ainda conforme (KOWATA, 2010), a transformação da sentença s_i em triplas no formato {conceito – relação – conceito}, para este trabalho designado de “tripla conceitual”, requer a identificação prévia dos elementos candidatos, por meio de reconhecimento de padrões linguísticos e de entidades nomeadas.

A construção de proposições a partir de conceitos e de relações é delimitada por processos decisórios de rearranjo entre os elementos do discurso, mapeados morfossintaticamente, com o objetivo de formular triplas conceituais (proposicionais). Em (KOWATA, 2010), são propostas sete atividades básicas para a construção de mapas conceituais a partir de texto em Português do Brasil. É relatado também uma experimentação da metodologia e os detalhes que a cercam. Para melhor entendimento, as etapas foram resumidas adiante:

i. O conteúdo do documento é normalizado, eliminando formatações impróprias ou o convertendo para apresentação textual;

ii. O texto é separado em orações (frases) ou em sentenças (conjunto de orações), utilizando-se da identificação de caracteres de pontuação e finalizadores. O desafio é a correta distinção entre sinais de fim de sentença e os elementos de demarcação (datas números, abreviações, etc.);

iii. As sentenças são divididas em entidades nomeadas (palavras, nomes, numerais, etc.), por meio de algoritmos de reconhecimento;

iv. Cada entidade é classificada (etiquetada) morfologicamente, ou seja, é verificado se corresponde a substantivo, verbo, pronome, preposição, advérbio, conjunção, artigo ou outra classe gramática específica;

v. Por meio de um conjunto de padrões linguísticos e pelas etiquetas morfológicas, grupos sintagmáticos são identificados: sintagmas nominais, sintagmas verbais, sintagmas preposicionais, sintagmas adjetivais, sintagmas adverbiais ou outros tipos definidos em expressões regulares e em autômatos finitos. Relacionam-se os sintagmas à identificação de elementos candidatos a conceitos e a relações;

vi. Uma vez estabelecidos os candidatos (sintagmas), estruturam-se como nós de grafos. Aqueles com núcleos verbais ou preposicionais são mapeados em arestas e os com núcleos nominais são mapeados em nós. Um interpretador de dependências pesquisa a posição mais adequada no grafo para subsumir novos elementos, por meio de proximidade dos nós afins, de acordo com regras de aproximação pré-definidas. Não são abordadas todas as circunstâncias sintáticas possíveis, mas as suficientes para contemplar a generalidade da proposta;

vii. Por último, os grafos são percorridos e arranjos na forma de proposições são definidos, gerando-se estruturas passíveis de se representar graficamente por mapa conceitual.

Em (KOWATA, 2010), os grafos foram escolhidos como estruturas intermediárias ao algoritmo. Do mesmo modo, este trabalho propõe a utilização dos grafos conceituais como etapa constituinte da estruturação de documentos voltados à disseminação seletiva, com a seguinte ressalva: ainda que as relações sejam visualizadas como arestas (úteis para mapas conceituais), o relacionamento quando tratado como vértice do grafo (Figura 3) transfere a solução outras possibilidades computacionais, que serão exploradas nas próximas seções.

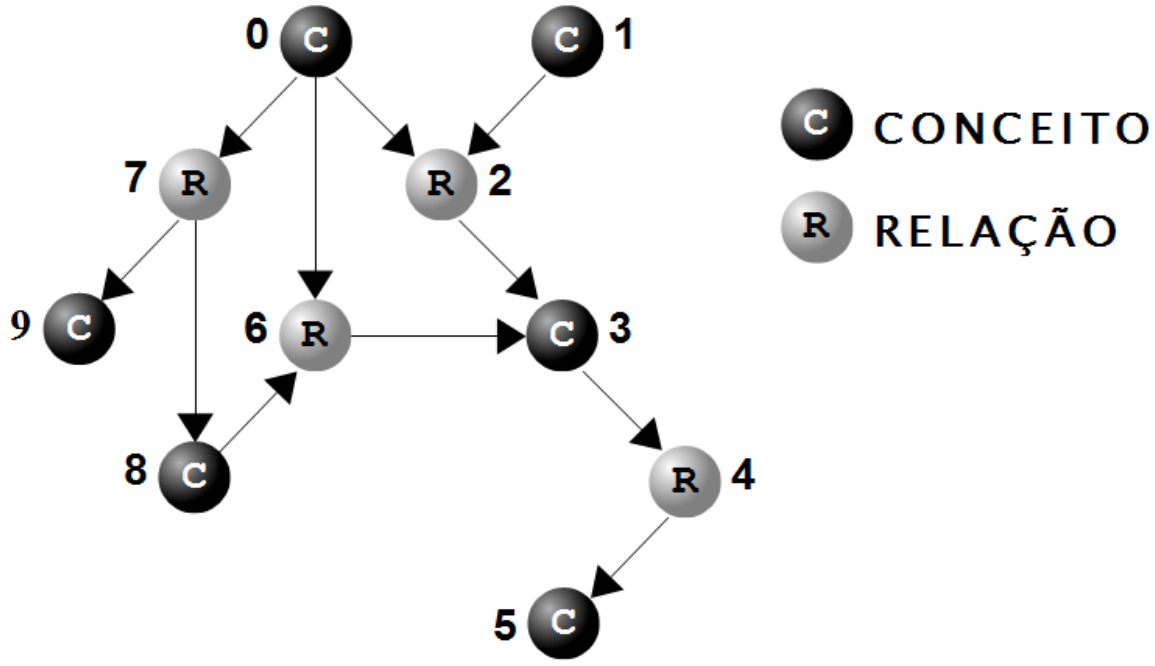

Figura 3. Ilustração de grafo conceitual

Por conseguinte, os grafos conceituais, para esta dissertação, serão a organização básica do conjunto de triplas conceituais detectadas a partir de textos, de forma a se apresentar como um tipo especial de grafo colorido direcionado, em que vértices conceitos terão como adjacentes somente vértices relação, e relação somente vértices conceitos.

Quando os vértices conceitos se portarem como filhos da relação, receberão denominação “predicado”, que na Figura 1, são enumerados em 3, 5, 8 e 9. De outra forma, quando os vértices conceitos forem pais, então serão denominados de “sujeito”, ilustrado na Figura 1 por 0, 1, 3, 8. Nota-se que os vértices 3 e 8 são híbridos, ora sujeito, ora predicado.

Outra condição especial proposta por este trabalho é que as triplas conceituais não se compunham somente da forma binária, ou seja, sejam revistas para {conceitos – relação – conceitos} ou, de acordo com o que foi proposto no parágrafo anterior, {sujeitos – relação – predicados}. Essa iniciativa visa retratar as seguintes práticas: otimização de memória, heurísticas para resolução de anáforas e estruturação do discurso baseado em elementos centrais (FREITAS, 2005).

O foco deste trabalho, quanto à geração de grafos conceituais a partir de texto, é adaptar o modelo de forma a servir como instrumento para estruturação de conteúdo e como diretriz para a visão lógica dos documentos de um modelo Vetorial estendido (ver Seção 4.2.3).

Para evidenciar as diferenças nas abordagens, seja o exemplo de grafo conceitual gerado (Figura 4) a partir do texto: "*O governador Mário Pereira, do Paraná, e o secretário da agricultura José Carlos Tibúrcio aproveitaram o palanque da Exposição Agropecuária de Londrina para cobrar o ministro da Agricultura, Synval Gazzeli, novo modelo de política agrícola capaz de estimular investimentos e o crescimento da produção.*"

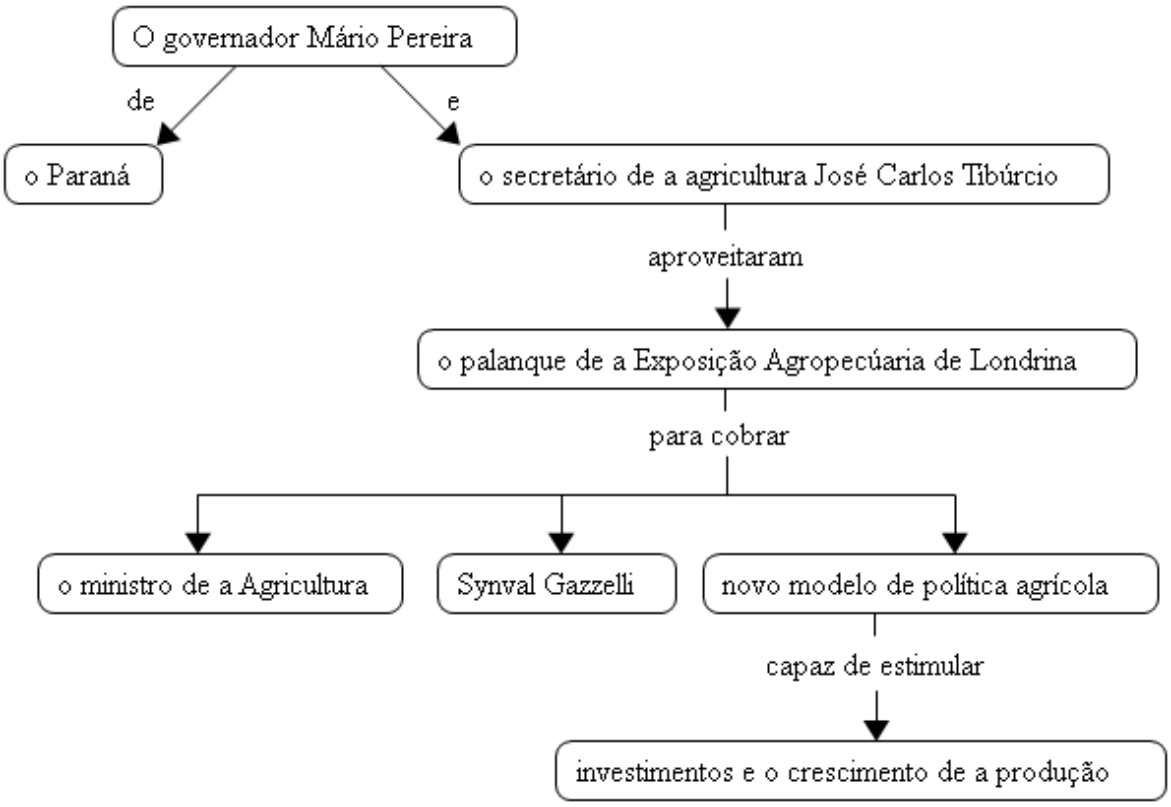

Figura 4. Um grafo conceitual a partir de texto – adaptado de (KOWATA, 2010)

Para (KOWATA, 2010), o aprofundamento de questões linguísticas, sobretudo nos mecanismos de tratamento de anáforas, número e gênero, entre outros, são necessários para o aperfeiçoamento da técnica. Também é relatado que a ausência de expressões regulares para tratar certos tipos de advérbios, conjunções e artigos condicionou a comportamentos indevidos, por exemplo, descartes de palavras.

Esta dissertação também propõe alguns questionamentos, como ilustrado na Figura 5, em relação a: (i) proposições com aparente igualdade quanto à estrutura morfossintática e quanto ao contexto possuem critérios diferentes na formação dos grafos conceituais; (ii) existem relações verbais não ligadas diretamente aos sujeitos da oração e (iii) existem relações sem valor computacional, para esta dissertação, principalmente para aquelas originadas de preposições, de pronomes relativos e de conjunções, apesar de essas estruturas serem válidas enquanto elemento de ligação.

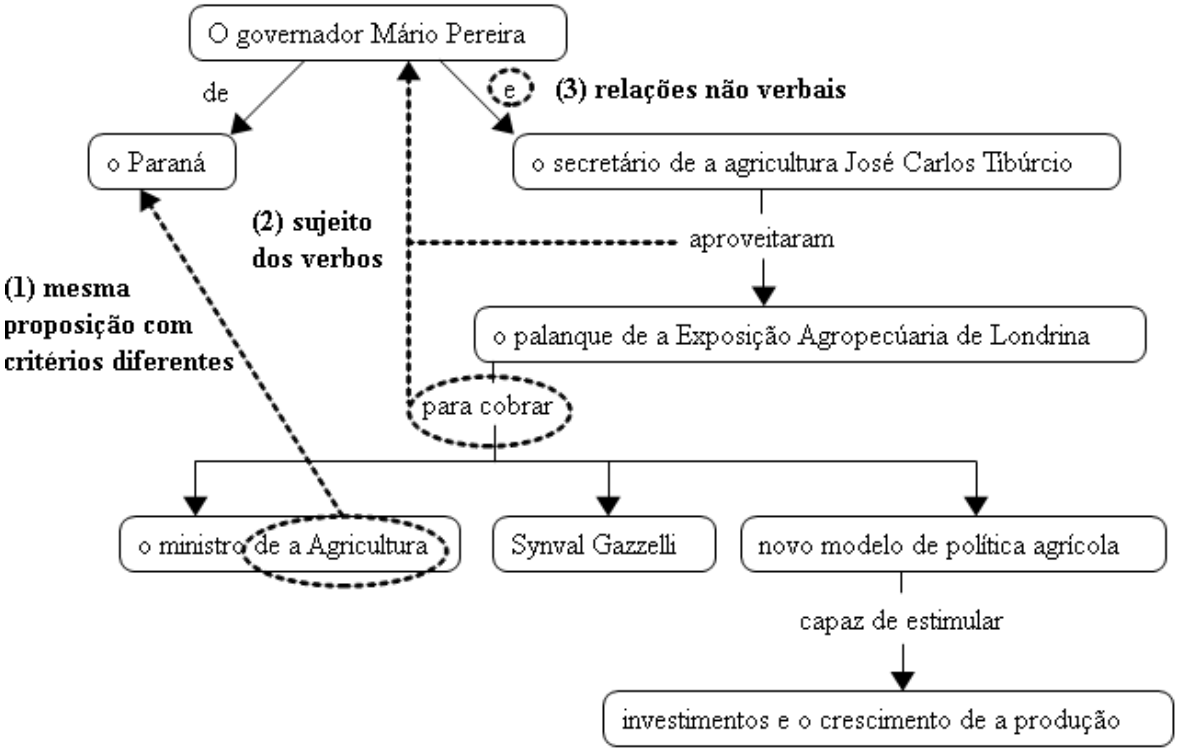

Figura 5. Questionamentos sobre grafo conceitual gerado em (KOWATA, 2010)

De acordo com (KOWATA, 2010), o uso de elementos de marcação (sinais de pontuação) como argumento de expressões regulares pode ser a causa de (i). À medida que há uma sobreposição de regras, em certas situações, há prevalência por regras de pontuação.

Em (ii), mesmo que indiretamente ligadas por vértices intermediários, se fosse mantida a relação entre sujeitos e verbos, a solução proveria melhor integridade quanto à semântica textual. O sentido inerente (correspondência com o texto) do grafo conceitual poderá ficar comprometido ou incompleto se, por algum motivo, um vértice (sujeito) for descartado.

De maneira semelhante, ao analisar isoladamente uma tripla formada por relações não verbais, essa estabelece pouco valor informacional ao conjunto, por exemplo, para a tripla {“O governador Mário Pereira”, “e”, “o secretário de a agricultura José Carlos Tibúrcio”} não se pode estabelecer uma afirmação conclusiva, além de ser difícil a distinção entre quem é o sujeito e quem é o predicado. Para este trabalho, optou-se por esses conceitos serem uma unidade, ou seja, um único vértice sujeito. Outra opção foi a de adotar somente relações verbais, estritamente originadas de sintagmas verbais, mostrado na Figura 6.

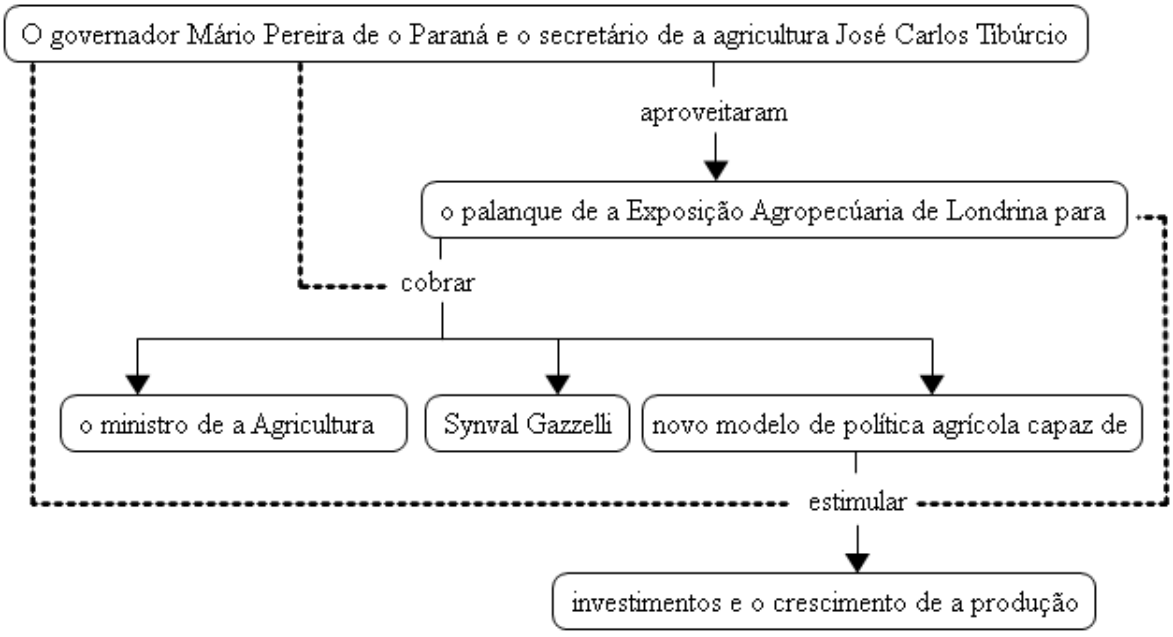

Figura 6. Proposta de grafo conceitual para esta dissertação

As atividades para geração desse grafo conceitual (Figura 6) serão as correspondentes em (KOWATA, 2010), desde que as condições expressas sejam observadas na composição das expressões regulares. Destaca-se um ponto importante relacionado à identificação de elementos centrais ao grafo: o vértice mais acima da última figura é aquele que possui o maior número de filhos (3) e o menor número de pais (0), ou seja, um indicativo de relevância. Em ordem inversa, vale-se para o vértice mais abaixo.

A característica manifestada no último parágrafo conduz a uma heurística voltada à seleção de elementos de valor informacional e à síntese desses: o grau dos vértices é um indicativo de fator de ponderação do grafo, em que vértices com maior número de filhos e menor quantidade de pais serão candidatos a sujeitos principais (maior ponderação) e o mesmo vale ao inverso, ou seja, vértices com maior número de pais e menor número de filhos serão candidatos a predicados principais (maior ponderação). Logo, a relação que interliga sujeitos principais a predicados principais deverá ter maior importância.

Seguindo esse raciocínio, no exemplo da Figura 6 a tripla {“O governador Mário Pereira e o secretário de a agricultura José Carlos Tibúrcio”, “aproveitaram”, “o palanque de a Exposição Agropecuária de Londrina”}, teria a menor relevância dentre as outras triplas, o que na prática se confirma, pois tal informação pode ser vista como algo adicional ao contexto, em que se mantém o sentido do grafo conceitual mesmo com exclusão da tripla. Para este trabalho, a utilidade dessa observação se manifestará na composição dos pesos vetoriais (ver Seção 4.2.3) e na resolução de anáforas (FREITAS, 2005).

A abordagem computacional proposta por (KOWATA, 2010) parte do reconhecimento de mapas conceituais a partir de um único documento (texto) e logo, provavelmente feito por um grupo reduzido de autores ou um único autor. Desse modo, pode-se afirmar que esse artefato está próximo da construção individualizada do conceito, estando presente apenas no modelo mental de cada indivíduo, o que pode tornar impreciso o compartilhamento ou a identificação de correlações semânticas (conceituais) entre discursos. De forma a superar tal fato, esta dissertação propõe o uso de arcabouços linguísticos, explicado na próxima seção.

4.2.2 Utilização de arcabouços linguísticos

Em sistemas de computação é comum o uso de mecanismos para correlacionar signos e aportar significantes computacionais, por meio de estruturas de representação ou de gerência do conhecimento, de comunidades científicas ou de grupo de especialistas. Em recente literatura, encontram-se aqueles sistemas apoiados por tesauros e por ontologias (MEDEIROS, 2011).

Em (MEDEIROS, 2011), encontram-se definições de vários autores sobre tesauro e sobre ontologia. Resumidamente, um tesauro é um sistema hierárquico e semântico baseado em conceitos e vocabulário controlado, apresentando relações entre os termos constituintes.

O termo “ontologia” é utilizado para designar várias correntes: disciplina filosófica, sistemas conceituais semânticos e meta-nível de teorias lógicas. Esta dissertação volta-se para as abordagens interessadas na construção de vocabulários terminológicos.

Em (LANCASTER, 2004), é exposto à evolução de padrões de construção de vocabulários controlados, evidenciando duas linhas: europeia, baseada em princípios classificatórios dos conceitos, e outra norte-americana, baseada na indexação de assuntos.

Nota-se que os sistemas europeus tendem a ontologias de fundamentação (GUIZZARDI, 2005), enquanto as norte-americanas aprofundam-se na composição de redes de palavras (MILLER, 1995).

A diferença principal dessas abordagens refere-se à formalização das relações entre os conceitos (LANCASTER, 2004), em que correntes europeias primam pela fundamentação, já as norte-americanas consideram implícitas aos sistemas de termos (notório à linguagem e consensual).

Não obstante a esse paralelismo metodológico, este trabalho considera que, mesmo nas duas linhas científicas, existe um potencial sistemático para, dado um conceito extraído de um texto e pertencente a um vértice do grafo conceitual, outro conjunto de termos pode ser retornado e esse conjunto corresponder a um sentido e a uma finalidade, que o discriminará dos demais conjuntos, em relação à palavra.

Para tanto, é proposto que esse conjunto deva ser proveniente de sistemas de conhecimento compartilhado. Para qualquer tipo de método com essa especificidade cunhou-se o nome de “arcabouço linguístico”, por iniciativa desta dissertação. Ainda que existam diferentes formas de se compor um arcabouço linguistico dentre as áreas de pesquisa, europeia ou norte-americana, futuros trabalhos deverão observar as diversas necessidades e cenários que possam existir (ver Seção 6.2). Como exemplo de arcabouço, este trabalho optou por mecanismos simples, de fácil aplicabilidade, cabendo a novos trabalhos a evolução desses.

Para exemplificar o arcabouço linguístico que será base para as próximas teorizações, optou-se por utilizar pesquisas correspondentes do Brasil para a linha norte-americana, dentre os quais se cita o trabalho de (DIAS-DA-SILVA, FELIPPO e NUNES, 2008) (DIAS-DA-SILVA, FELIPPO e NUNES, 2008), que relata uma rede lexical semântica, para termos do Português do Brasil, capaz de indexar sinônimos, antônimos, hiperônimos, hipônimos, merônimos, holônimos entre outros tipos de relações, inerentes a linguagem.

Pode-se formalizar um arcabouço linguistico por teoria dos conjuntos: existe um conjunto *K*, tal que *K* é o domínio e contradomínio de todos os signos, ou a base de conhecimento, e existem funções do tipo $f: K \rightarrow K$, que associam um signo a outros, mesclando os significados. Nesta dissertação, foram definidas as seguintes funções:

i. *i(k)* – função identidade: dado um signo *k* (termo, palavra, sintagma, etc.), são retornados aqueles signos em que todas as afirmações válidas para *k* também são válidas para esses. Por exemplo, podem-se adotar os sinônimos e os hipônimos de uma rede lexical semântica – *i*(‘feliz’) = {‘feliz’, ‘alegre’, ‘contente’};

ii. *s(k)* – função similaridade: dado um signo *k*, são retornados aqueles signos em que parte das afirmações válidas para *k* também são válidas para esses. Por exemplo, podem-se adotar as derivações, hiperônimos, merônimos, holônimos de uma rede lexical semântica – *s*(‘feliz’) = {‘ditoso’};

iii. *c(k)* – função contradição: dado um signo *k*, são retornados aqueles signos em que todas as afirmações válidas para *k* não são válidas para esses. Por exemplo, podem-se adotar os antônimos de uma rede lexical semântica – *c*(‘feliz’) = {‘infeliz’};

Desse modo, pode-se afirmar que as funções identidade e contradição são sobrejetivas, visto que, na função identidade, o argumento sempre estará no contradomínio e, na função contradição, a negação do argumento. Entende-se por negação o signo correspondente a forma negativa desse, geralmente, acrescido da palavra “não” (“não feliz”) ou do prefixo “in” (“infeliz”) e de outros.

4.2.3 Indexação de triplas conceituais

Segundo (LANCASTER, 2004), a indexação de conteúdo se apresenta de duas formas:

i. Indexação seletiva: propõe a generalização do documento em classes abstratas de organização;

ii. Indexação exaustiva: proporciona indicações mais específicas, possibilitando maior número de pontos de acesso.

O propósito da indexação de triplas conceituais é rever o modelo Vetorial, de maneira que se consiga identificar automaticamente correlações semânticas, ou relações CST, entre os discursos de cada novo documento, por meio de indexação exaustiva e de busca vetorial a base indexada.

Dado que a tripla representa proposições sobre conceitos, pode-se utilizar de expansão da indexação, por meio dos arcabouços linguísticos, para reafirmar a proposição, de forma que a reescrita corresponda ao que se espera da correlação.

A intenção é que, gerando-se o grafo conceitual de novos textos, as triplas conceituais sejam representadas de forma vetorial contemplando possíveis correlações com triplas de outros documentos (ver Seção 4.2.4).

Por exemplo, seja a tripla conceitual {“Mário”, “investe”, “agricultura”}, ela poderá ser indexada também na forma {“Mário”, “incentiva”, “agricultura”}, para a correlação identidade.

Assim, um documento não seria indexado como um único vetor de *n* dimensões (modelo Vetorial clássico), mas como um conjunto de vetores índices correspondentes as triplas encontradas no conteúdo, multiplicado pela quantidade de correlações abordadas.

Sujeitos, relação e predicados formarão espaços vetoriais diferentes, respectivamente. Ou seja, o espaço vetorial clássico (SALTON, WONG e YANG, 1975) será subdivido em três espaços: espaço dos sujeitos, espaço das relações e espaço dos predicados.

A identificação das correlações se estabelece quando há correspondência entre os vetores nos três espaços, ou, no campo vetorial (conjunto de espaços). Assim, a similaridade entre triplas pode ser calculada pela multiplicação dos cossenos.

$$sim(T_i, T_j) = cos(\overrightarrow{S_i}, \overrightarrow{S_j}) \cdot cos(\overrightarrow{R_i}, \overrightarrow{R_j}) \cdot cos(\overrightarrow{P_i}, \overrightarrow{P_j})$$

$\overrightarrow{S_n}$ = vetor dos termos dos sujeitos da tripla *n*

$\overrightarrow{R_n}$ = vetor dos termos da relação da tripla *n*

$\overrightarrow{P_n}$ = vetor dos termos predicados da tripla *n*

Dessa forma, as triplas serão indexadas no campo vetorial, formado por três espaços, e os espaços terão a dimensão correspondente aos termos indexados de cada tripla, como ilustrado na Tabela 6. O valor da similaridade entre as triplas será afetado por cada um dos espaços.

Tabela 6. Exemplo simples de indexação de triplas conceituais

D1 - "Paraná solicitou bolsa agrícola"
D2 - "Paraná investe na bolsa"

Termos	D1			D2		
	S1	**R1**	**P1**	**S2**	**R2**	**P2**
"Paraná"	1			1		
"bolsa"			1			1
"solicitou"		1			0	
"agrícola"			1			0
"investe"		0			1	

$$cos(\overrightarrow{S_i}, \overrightarrow{S_j}) \cdot cos(\overrightarrow{R_i}, \overrightarrow{R_j}) \cdot cos(\overrightarrow{P_i}, \overrightarrow{P_j})$$

$$sim(T_i, T_j) = 1 \cdot 0 \cdot 0.7 = 0$$

No exemplo da Tabela 6, a comparação entre os documentos D1 e D2 assume valor zero, visto que os vetores não possuem semelhança no espaço relação, mesmo que em outros espaços sejam próximos. Nota-se que houve uma evolução do modelo Vetorial quanto à polissemia.

Para os pesos dos vetores, pode ser adotada a heurística expressa anteriormente, ou seja, a quantidade (grau) de filhos, quando um vértice for sujeito, e a quantidade de pais, quando for predicado, ou funções matemáticas que normalizem tais características.

O mesmo vale para as relações, cuja relevância deverá refletir o fato de que se operam sobre conceitos que são menos ou mais significativos para o grafo. Um estudo mais aprofundado sobre os pesos dos vetores é proposto como trabalho futuro (ver Seção 6.2).

4.2.4 Correlações intertextuais básicas

A partir do refinamento e da reclassificação das relações CST, exposta em (JORGE, 2010), este trabalho optou por abordar somente o conjunto de relações intertextuais básicas em que são definidas correspondentes no arcabouço linguístico.

Desse modo, enquanto o arcabouço linguistico fornece um mecanismo para correlacionar signos, o modelo Vetorial estendido será o expoente do arcabouço para triplas conceituais, que representam as proposições de um documento.

Utiliza-se das funções presentes no arcabouço linguístico para delimitar o espaço solução do modelo de indexação das triplas conceituais de maneira que, dado uma tripla conceitual, possa se estabelecer um conjunto de outras triplas (vetores) que possuem um significante, agora proposicional.

De certo modo, pode-se estabelecer uma referência à lógica proposicional para elucidar como proposições podem ser reescritas na forma de identidade, parcialidade e negação (contradição). Porém, este trabalho optou por não se utilizar de sistemas formais de cálculo proposicional, de regras de derivação ou de modelos axiomáticos para formalizar a solução.

De fato, a proposta é se utilizar, ainda que intuitivamente, dos conceitos da lógica proposicional. Porém, ao referir-se como solução vetorial, atentou-se que a junção dessas perspectivas poderá ser trabalho de aprofundamento teórico no futuro (ver Seção 6.2), uma vez que não é o objetivo desta dissertação.

Outrora designadas como função, em arcabouços linguísticos, pode-se refinar as relações CST para as seguintes correlações semânticas: identidade, similaridade e contradição (que para o modelo serão as formas de indexação das triplas conceituais). Uma formalização é apresentada:

i. Identidade – correlação entre triplas conceituais em que, dada uma tripla conceitual $\{S, R, P\}$, tal que S é o conjunto de todos os sujeitos da relação R e que P é o conjunto de todos os predicados, têm-se os seguintes vetores a serem indexados:

$$\left\{\overrightarrow{\imath(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{\imath(P)}\right\},\ i(X) \therefore signos\ de\ X$$

Ou seja, aplica-se $i(k)$ (função identidade) para todos os signos de S, R e P.

$$\left\{\overrightarrow{\imath(S)},\ \overrightarrow{c(R)},\ \overrightarrow{c(P)}\right\},\ \left\{\overrightarrow{c(S)},\ \overrightarrow{c(R)},\ \overrightarrow{\imath(P)}\right\},\ \left\{\overrightarrow{c(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{c(P)}\right\}$$

Para as orações do tipo: “Mário é feliz” e “Mário não é triste”.

ii. Similaridade – do mesmo raciocínio do item anterior:

$$\left\{\overrightarrow{\imath(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{s(P)}\right\},\ \left\{\overrightarrow{\imath(S)},\ \overrightarrow{s(R)},\ \overrightarrow{s(P)}\right\}$$

$$\left\{\overrightarrow{s(S)},\ \overrightarrow{s(R)},\ \overrightarrow{s(P)}\right\},\left\{\overrightarrow{s(S)},\ \overrightarrow{s(R)},\ \overrightarrow{\imath(P)}\right\}$$

$$\left\{\overrightarrow{s(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{\imath(P)}\right\},\left\{\overrightarrow{s(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{s(P)}\right\}$$

iii. Contradição – do mesmo raciocínio do item anterior:

$$\left\{\overrightarrow{c(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{\imath(P)}\right\},\ \left\{\overrightarrow{\imath(S)},\ \overrightarrow{c(R)},\ \overrightarrow{\imath(P)}\right\}$$

$$\left\{\overrightarrow{\imath(S)},\ \overrightarrow{\imath(R)},\ \overrightarrow{c(P)}\right\},\ \left\{\overrightarrow{c(S)},\ \overrightarrow{c(R)},\ \overrightarrow{c(P)}\right\}$$

O procedimento para indexação de grafos conceituais é uma atividade que consiste na expansão dos signos de cada tripla conceitual (sub-grafo), por meio de funções do arcabouço linguistico, e na geração de vetores para cada espaço do campo vetorial (sujeito, relação e predicado), como ilustrado na Figura 7.

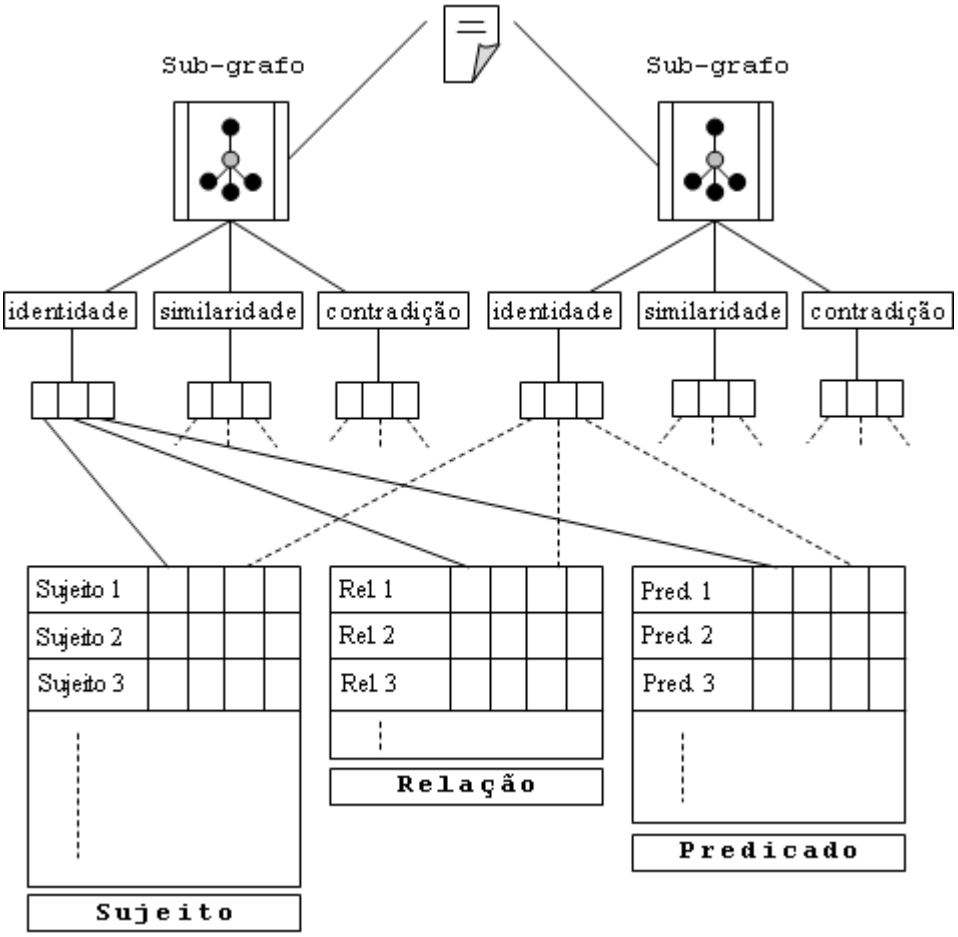

Figura 7. Indexação das correlações intertextuais básicas

Observa-se na Figura 7 que as dimensões entre os espaços podem ser diferentes, o que na prática é provável que se concretize, visto que o número de verbos é menor que o número de substantivos.

A identificação das correlações entre textos pode ser realizada quando, dado um novo documento e gerado o grafo conceitual, para cada tripla formada realiza-se uma consulta vetorial a coleção de triplas expandidas já indexadas. O resultado pode ser guardado em listas ordenadas pelo valor de similaridade (cosseno) entre as triplas, como mostrado na Figura 8.

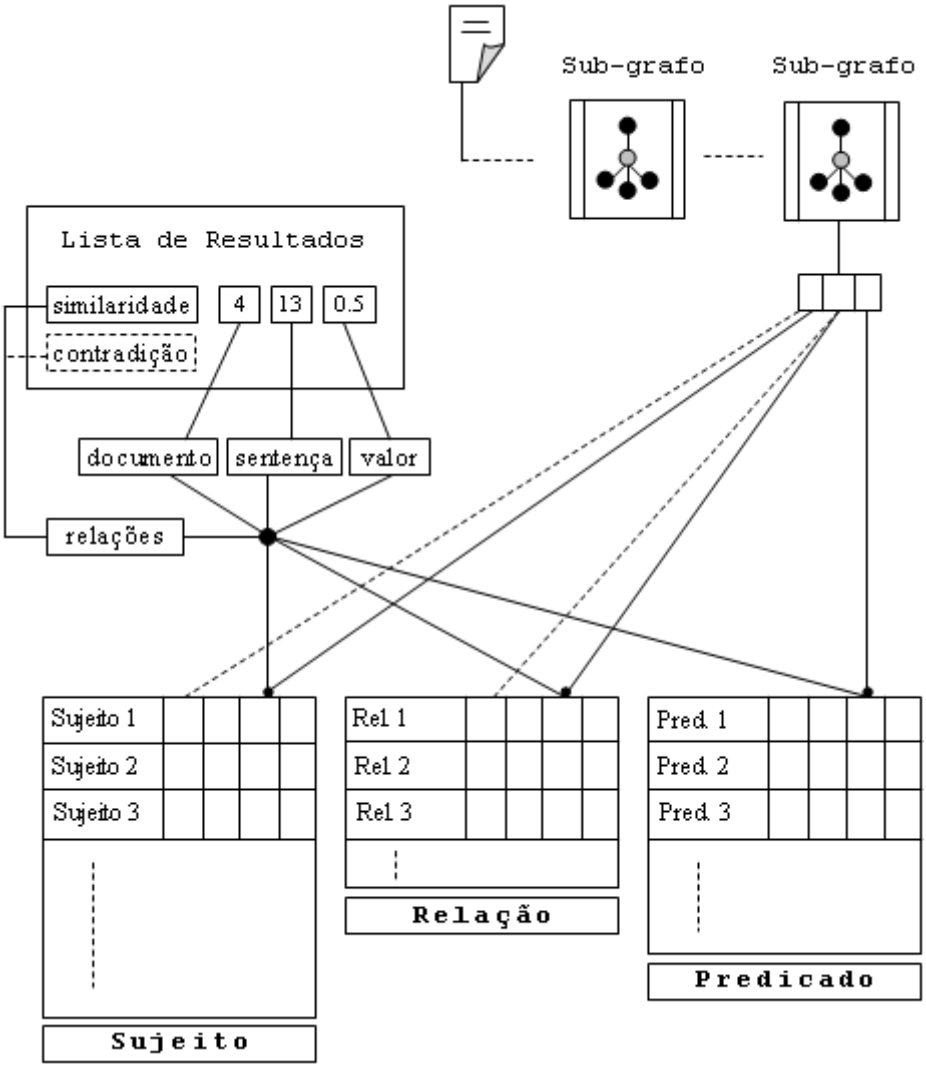

Figura 8. Pesquisa das correlações semânticas

Ressalta-se que o modelo proposto é útil para identificação de correlações semânticas entre textos. Para recuperação de informação, não se pode presumir que sempre orações serão usadas como parâmetro de pesquisa. Já para sistemas de disseminação seletiva, o modelo é adequado à medida que o interesse possa ser expresso textualmente, o que é usual.

É evidente que o processo de identificação das correlações dependerá da eficácia dos sistemas condicionantes da proposta (gerador de grafo conceitual e arcabouço linguístico). O importante é que esses são passíveis de construção, como mostrado no Capítulo 5.

Ao passo que novos documentos forem sendo indexados e as relações estabelecidas, a base de correlações se tornará uma rede semântica entre documentos, evidenciado na Figura 9.

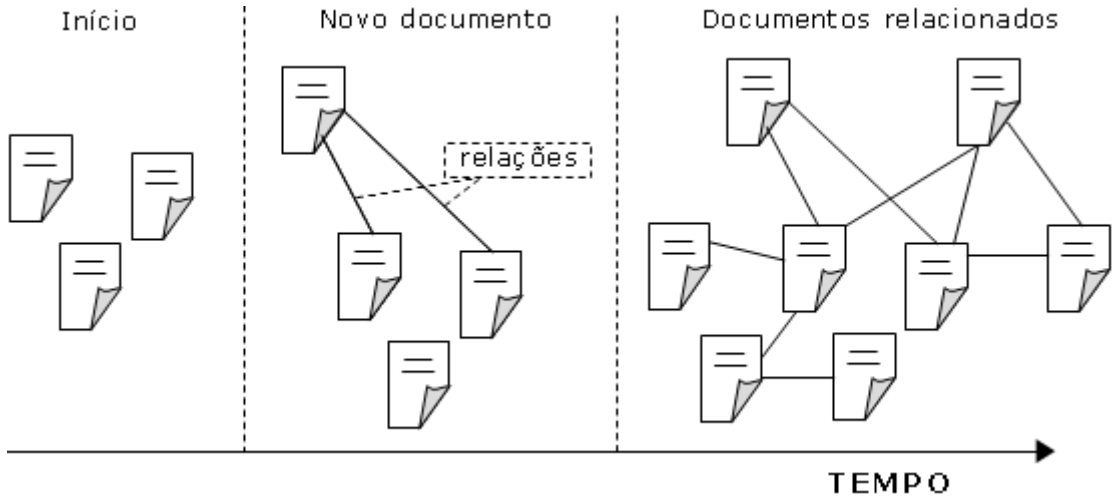

Figura 9. Evolução da rede semântica entre documentos

A prospecção para modelos associativos em rede (AGOSTI e MARCHETTI, 1992) é um dos pontos a serem trabalhados e novas propostas poderão investigar o uso estatístico da rede semântica para validar e invalidar correlações (ver Seção 6.2), ainda que a rede semântica seja construída de forma automática.

Outras correlações semânticas, além das básicas mencionadas, podem ser propostas, desde que suportadas por arcabouço linguístico. Por exemplo, se houvesse a função denominada temporalidade *t* que dado um signo *k* retornasse todos aqueles que possuem afirmações temporais sobre *k* – como em *t*(‘namoro’) = {‘noivado’, ‘casamento’, ‘separação’} – então se pode propor um correspondente para correlação semântica entre triplas conceituais.

Um aspecto a ser investigado na proposta são as orações na ordem invertida, por exemplo, “Feliz é Mário” e “Uma bolsa agrícola foi solicitada por Mário”. Nesses exemplos, a inversão não ocasionou mudança de sentido da frase, dado que as formas verbais permitem tal condição. Porém, em outras frases como “Agricultura incentiva Mário”, a coerência da oração foi comprometida e a inversão altera o sentido.

Pode-se optar pelo detrimento da última forma verbal, ainda que diminua a eficiência, e indexar também as triplas com o sujeito e o predicado invertidos, no entanto, diminuindo o peso desses vetores. Dessa forma, o resultado da comparação assumiria um valor menor do que a forma direta.

Resumidamente, o modelo Vetorial estendido proposto por esta dissertação pode ser visto como uma sequência de atividades voltadas à obtenção de correlações entre textos por meio de pesquisa vetorial a base de triplas conceituais estendidas e indexadas de forma exaustiva. Os passos sequenciais do modelo podem ser enumerados:

i. Para cada novo texto, gera-se o grafo conceitual, conforme o que foi detalhado nos parágrafos anteriores;

ii. Anterior a indexação do novo texto, as triplas proposicionais do grafo conceitual geram consultas vetoriais a base indexada, que possui informações suficientes para definir se o resultado corresponde às correlações de identidade, de similaridade ou de contradição (e outras possíveis);

iii. Pesquisada as correlações, parte-se para indexação do grafo. As triplas conceituais contidas no grafo são expandidas por meio das funções providas por arcabouço linguistico e definidas por cada correlação (identidade, similaridade e contradição);

iv. Toda tripla é indexada na forma de um conjunto de três vetores correspondentes ao campo vetorial formados pelos espaços sujeito, relação e predicado.

4.3 Considerações parciais

O algoritmo proposto por esta dissertação visa ser uma proposta voltada à classificação automática de correlações semânticas entre discursos. Devido ao estágio inicial, o modelo necessita de revisões, principalmente quanto à formalização. Avaliações mais criteriosas sobre a complexidade algorítmica também se fazem necessárias.

Não obstante a essas constatações, o próximo capítulo realiza experimentações do modelo, a partir da construção de ferramental, que segue as diretrizes exposta por essa metodologia. Os artefatos são utilizados na composição de um sistema de disseminação seletiva, que é o objetivo principal dessa dissertação.

5 ESTUDO DE CASOS

Com base na solução apresentada no capítulo anterior, o objetivo deste capítulo é realizar experimentações das propostas a fim de verificar a viabilidade dessas para sistemas de disseminação seletiva da informação.

O elemento central das proposições refere-se à possibilidade de identificar o interesse por meio das próprias produções, em que o receptor do sistema de disseminação também será o produtor – o que é nomeado por (PRIMO, 2007) de agente.

Para a etapa de seleção de conteúdo, pode-se adotar como critério a identificação de correlações semânticas entre textos de diferentes agentes. Desse modo, volta-se o trabalho a concepção de sistemas em que a representação do interesse é intuitiva, visto que remete a interação contínua por meio da produção de textos.

Como forma de assegurar que o modelo Vetorial estendido do Capítulo 4 pode ser usado para identificar correlações semânticas entre textos, detalha-se na Seção 5.1 um projeto de modelo Vetorial baseado no que fora proposto. Paralelamente, interfaces de visualização dos conceitos são apresentadas.

Na Seção 5.2, é elaborado outro projeto, agora voltado à criação de um sistema de disseminação seletiva da informação, tendo como suporte ferramental o projeto anterior. O principal desafio nessa composição é criar um sistema que se atente aos requisitos de representação de conteúdo, representação de interesse, formas de seleção, qualidade, veracidade, síntese, interface de mediação, interoperabilidade, redução do esforço cognitivo, facilidade de interação e construção do conhecimento, citados no Capítulo 2.

São propostas também, na Seção 5.3, algumas aplicações do sistema, em diferentes cenários, principalmente para sistemas colaborativos e para sistemas voltados a informática na educação, em que são comumente encontrados. Por fim, o sistema também é avaliado subjetivamente.

No final do capítulo, são discutidas considerações finais sobre a abordagem apresentada, em que são relatados méritos, necessidades e deficiências.

5.1 Experimentação do modelo Vetorial proposto

Objetivando a construção de um protótipo que contemple o modelo do Capítulo 4, são detalhados nas próximas seções os componentes de sistemas utilizados, tanto para a construção de uma ferramenta para indexação e pesquisa vetorial, quanto para os pré-requisitos funcionais, ou seja, um módulo de geração automática de grafos conceituais e outro módulo como arcabouço linguistico.

Por escolhas arquiteturais, o protótipo feito em (KOWATA, 2010) não foi reutilizado. Essa decisão reflete o fato de que as adaptações das regras e das expressões para geração do grafo implicariam em manutenção no código. Optou-se, então, por construir outro sistema que apoiasse a solução deste trabalho de forma prática. No entanto, a maior parte da técnica do protótipo original foi mantida.

5.1.1 Visão geral da solução

A criação do protótipo do modelo Vetorial iniciou-se pela pesquisa de bibliotecas que pudessem ser reutilizadas para o propósito do modelo. Os objetos foram organizados na forma de três componentes de sistemas, com finalidades específicas que respondessem as etapas da metodologia. Portanto, pode-se dividir o protótipo em três módulos principais: gerador de grafo conceitual, arcabouço linguístico e módulo vetorial.

No gerador de grafo conceitual foi utilizado o analisador morfossintático derivado do sistema CoGrOO (KINOSHITA, SALVADOR e MENEZES, 2007). O sistema é escrito na linguagem Java[2] e essa foi escolhida para desenvolvimento deste e dos outros módulos.

A ferramenta CoGrOO possui um conjunto de rotinas reutilizáveis, dentre as quais podem ser aproveitadas para as atividades de geração automática de mapas conceituais (e grafos): normalização de textos, identificação de sentenças, anotação morfossintática. A partir dessas fases, reproduziu-se o algoritmo de (KOWATA, 2010), fazendo-se os ajustes necessários já expostos.

[2] www.java.com/

Para validar os grafos gerados por meio da ferramenta, foi elaborada uma interface Web capaz de exibir e manipular o grafo. A interface pode ser conferida na Figura 10 a seguir. Tal mecanismo elaborado poderá ser fruto de aprofundamento em futuros trabalhos (ver Seção 6.2).

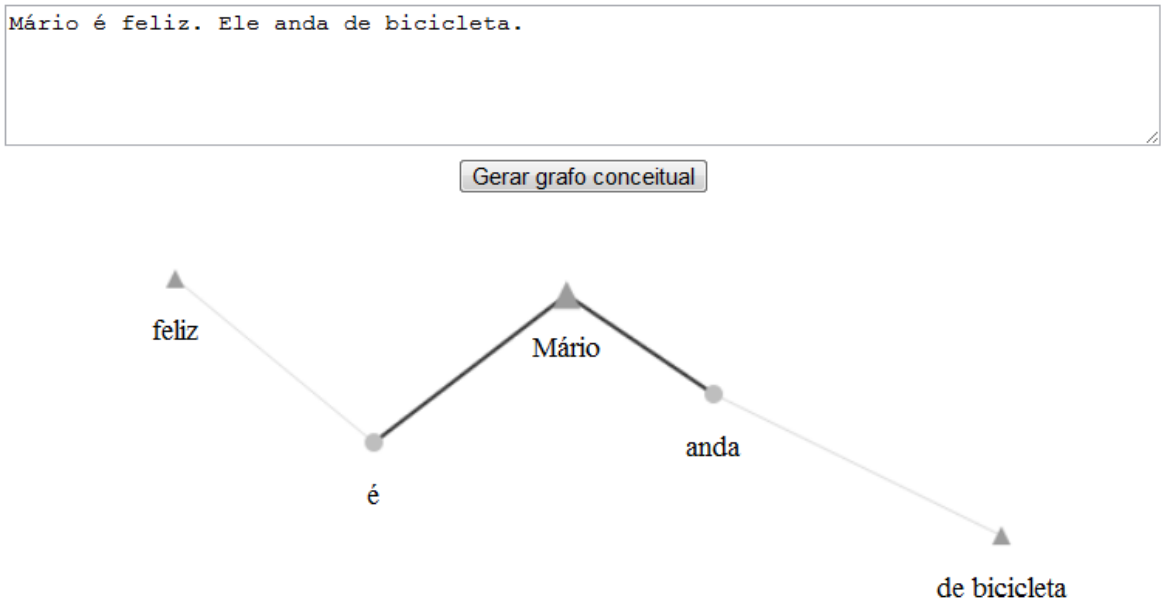

Figura 10. Interface para validação dos grafos conceituais

Nota-se na figura que a ferramenta foi capaz de resolver a anáfora pronominal “Ele anda de bicicleta”. A heurística utilizada para resolvê-la tem como base as teorias explicitadas em (FREITAS, 2005), que para este trabalho segue as premissas: o vértice sujeito anterior com o maior número de filhos terá a probabilidade maior de ser o sujeito de predicados anafóricos. No protótipo só foram tratados os pronominais. Essa técnica poderá ser fruto de aprofundamento em futuros trabalhos (ver Seção 6.2).

Para o segundo módulo do protótipo, referente à criação do arcabouço linguístico, foram utilizadas as bases de tesauros disponíveis para o Português do Brasil (DIAS-DA-SILVA, FELIPPO e NUNES, 2008) (DIAS-DA-SILVA, FELIPPO e NUNES, 2008). A dificuldade desse módulo passa-se pela leitura dos arquivos indexados, pelo entendimento dos formatos e pela necessidade de correspondência das estruturas morfossintáticas contidas no primeiro módulo.

De fato, os tesauros encontrados se utilizam somente das classes gramaticais substantivo, adjetivo, advérbio e verbo, geralmente nas formas masculina, singular e atemporal. Dado que a biblioteca CoGrOO é capaz de estimar e extrair o lema de muitas das palavras do Português do Brasil, foi utilizado esse artifício como maneira de não limitar a atuação do protótipo. Outra interface Web, para validação do arcabouço, também foi construída, como ilustrado na Figura 11. Tal interface elaborada poderá ser fruto de aprofundamento em futuros trabalhos (ver Seção 6.2). No exemplo, duas classes foram identificadas, advérbio e substantivo, além das relações de sinônimo, antônimo, derivado e similar. O sistema é capaz de identificar também hiperônimos, hipônimos, merônimos e holônimos.

Figura 11. Interface para validação do arcabouço linguístico

Por último, a biblioteca Apache Lucene [3] foi utilizada como ferramental tecnológico para manipular vetores. Disponível em código aberto, sob o domínio da Apache Foundation[4], foi inicialmente escrita em Java, mas há versões em diversas outras linguagens, como Delphi, Perl, C#, C++, Python, Ruby e PHP. O Lucene oferece níveis de abstração das técnicas do modelo Vetorial. A ferramenta provê também mecanismo de processamento distribuído, além de códigos para operação sobre textos. O Lucene também é proferido em vários projetos acadêmicos, por exemplo, em (LAURENCE, HIRSCH e SAEEDI, 2007).

Outra interface Web foi construída, a fim de verificar aplicabilidade da ferramenta, ilustrada na Figura 12. Tal proposta de interface poderá ser fruto de aprofundamento em futuros trabalhos (ver Seção 6.2).

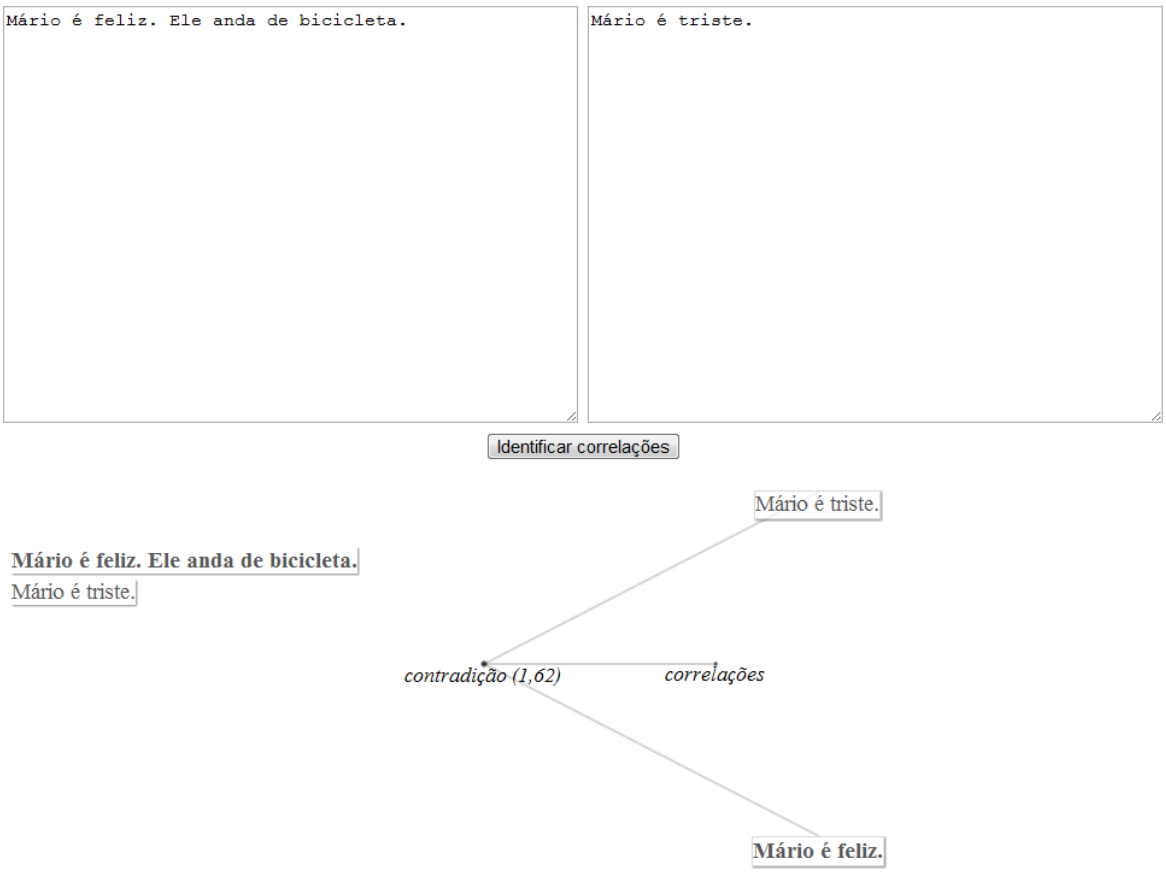

Figura 12. Interface para validação do módulo vetorial

[3] http://lucene.apache.org/

[4] http://apache.org/

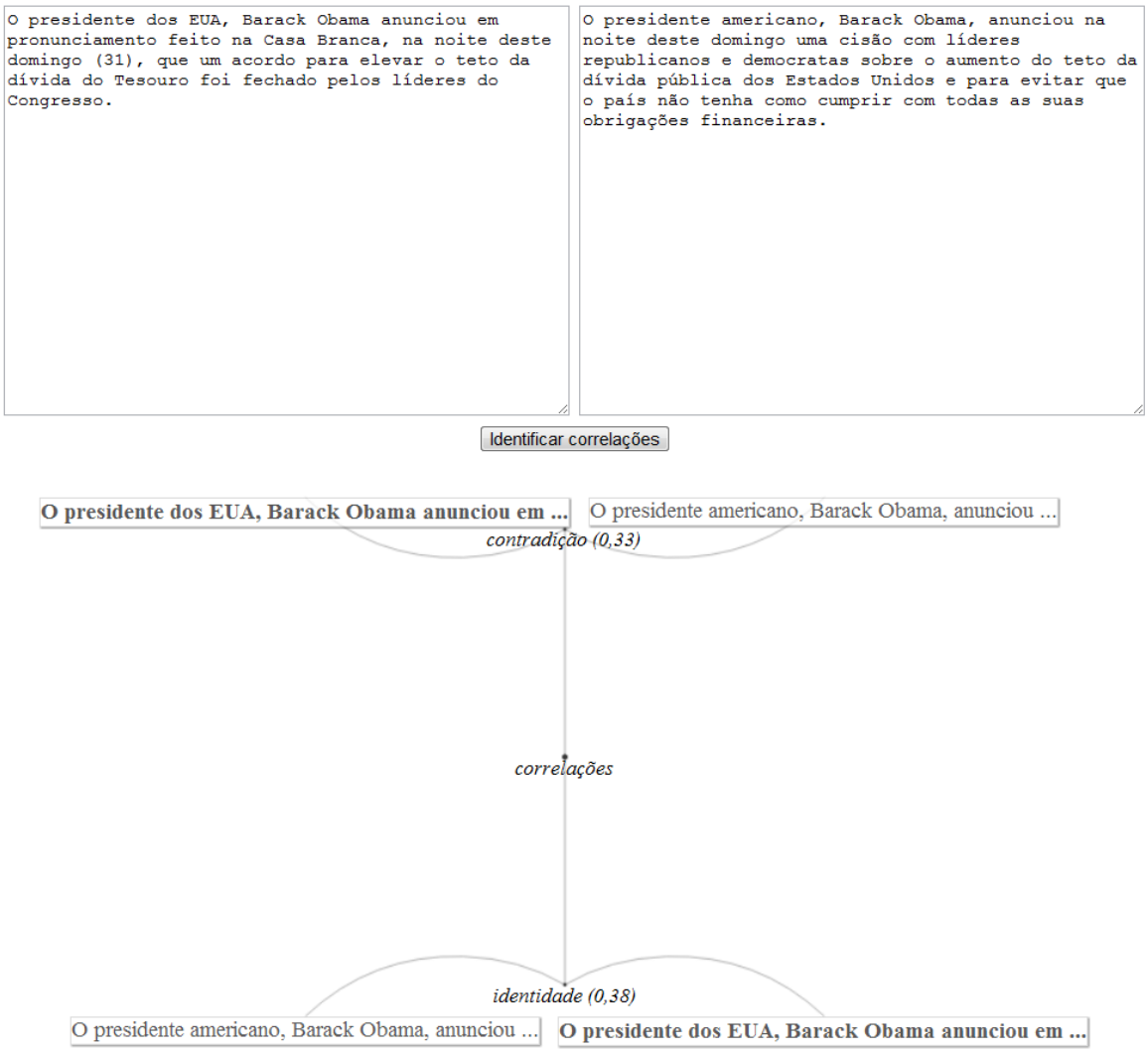

Figura 13. Exemplo de texto complexo tratado pelo modelo

No exemplo da Figura 13, duas correlações de pesos diferentes foram identificadas, retratando o fato de haver tanto a afirmativa idêntica, em "O presidente dos EUA, Barack Obama anunciou em pronunciamento feito na Casa Branca, na noite deste domingo" e "O presidente americano, Barack Obama anunciou na noite deste domingo"; como também o fato de haver uma oração contraditória, em "O presidente dos EUA, Barack Obama anunciou em pronunciamento feito na Casa Branca, na noite deste domingo (31) que um acordo para elevar o teto da dívida do Tesouro" e "O presidente americano, Barack Obama anunciou na noite deste domingo, uma cisão com os líderes republicanos e democratas sobre o aumento do teto da dívida pública".

5.1.2 Métricas de avaliação

As métricas de avaliação do modelo serão as mesmas que comumente são apropriadas em sistemas de recuperação da informação. A medição será feita por meio do comparativo entre as relações anotadas automaticamente e as anotadas de forma manual (ver Seção 5.1.3). As métricas escolhidas foram:

i. Precisão (*Precision*): proporção entre o número de documentos relevantes classificados (somente os que deveriam ser classificados, dentre todos os que foram anotados) e o número total de documentos anotados;

ii. Abrangência (*Recall*): proporção entre o número de documentos relevantes classificados (somente os que deveriam ser classificados, dentre todos os que foram anotados) e o número de documentos relevantes para o universo da classificação;

5.1.3 Descrição do ambiente

Para a realização dos experimentos voltados à análise quantitativa, foi utilizado um *corpus* composto por cinquenta coleções de textos jornalísticos, escritos em Português do Brasil, denominado *CSTNews* (ALEIXO e PARDO, 2008).

Cada coleção agrupa textos sobre os mesmos tópicos. O *corpus* foi anotado com relações CST por linguistas previamente treinados, obtendo resultados de concordância satisfatórios (ALEIXO e PARDO, 2008). No total, a base possui 195 documentos e 3.534 sentenças.

5.1.4 Resultados dos experimentos

O processamento das amostras iniciou-se pelo módulo de geração do grafo conceitual. O sistema foi capaz de identificar 10.371 conceitos e 4.908 relações entre conceitos, em média de 16,43 vértices por documento, e em média de 5,27 relações e 11,15 sujeitos, ou seja, 2,11 conceitos por relação. O número de sujeitos foi em média de 3,34. Já o número de predicados de 4,62 por tripla conceitual.

Os registros das relações anotadas manualmente foram estruturados para compor uma tabela de correlações válidas. Para essa análise, a correlação “similaridade” não foi considerada, devido ao fato de não haver correspondente nas relações CST definidas em (JORGE, 2010). Portanto, foram consideradas para a análise somente as relações identidade e contradição.

A cada correlação identificada automaticamente pelo protótipo, buscou-se na tabela se a relação era válida, ou seja, se também foi anotada manualmente. Essa informação foi tratada no sistema de avaliação e ao fim do processo, as precisões e a abrangências das coleções foram calculadas. Os resultados estão expressos na Tabela 7 e Figura 14.

Tabela 7. Resultados da avaliação do protótipo do modelo Vetorial estendido

Identidade		**Contradição**	
Precisão	**Abrangência**	**Precisão**	**Abrangência**
0,97380	0,039216	0,692246	0,033333
0,933887	0,058824	0,523664	0,046667
0,796846	0,078431	0,346123	0,060000
0,659062	0,098039	0,267797	0,073333
0,583679	0,117647	0,230749	0,086667
0,534833	0,137255	0,208387	0,100000
0,501567	0,156863	0,193097	0,113333
0,473579	0,176471	0,181818	0,126667
0,452657	0,196078	0,173062	0,140000
0,435558	0,215686	0,166009	0,153333
0,421240	0,235294	0,160171	0,166667
0,409018	0,254902	0,155232	0,180000
0,398423	0,274510	0,150982	0,193333
0,389120	0,294118	0,147273	0,206667
0,380862	0,313725	0,143997	0,220000
0,373465	0,333333	0,141076	0,233333
0,366787	0,352941	0,138449	0,246667
0,360717	0,372549	0,136069	0,260000
0,355166	0,392157	0,133899	0,273333
0,350062	0,411765	0,131909	0,286667
0,345348	0,431373	0,130074	0,300000
0,340975	0,450980	0,128376	0,313333
0,336903	0,470588	0,126798	0,326667
0,333098	0,490196	0,125326	0,340000
0,329531	0,509804	0,123948	0,353333
0,326178	0,529412	0,122655	0,366667
0,323018	0,549020	0,121437	0,380000
0,320032	0,568627	0,120288	0,393333
0,317205	0,588235	0,119202	0,406667

0,314522	0,607843	0,118171	0,420000
0,311971	0,627451	0,117193	0,433333
0,309542	0,647059	0,116262	0,446667
0,307225	0,666667	0,115374	0,460000
0,305010	0,686275	0,114527	0,473333
0,302891	0,705882	0,113717	0,486667
0,300861	0,725490	0,112941	0,500000
0,298913	0,745098	0,112197	0,513333
0,297042	0,764706	0,111483	0,526667
0,295242	0,784314	0,110796	0,540000
0,293510	0,803922	0,110136	0,553333
0,291840	0,823529	0,109499	0,566667
0,290229	0,843137	0,108886	0,580000
0,288673	0,862745	0,108293	0,593333
0,287170	0,882353	0,107721	0,606667
0,285716	0,901961	0,107168	0,620000
0,284308	0,921569	0,106633	0,633333
0,282945	0,941176	0,106115	0,646667
0,281623	0,960784	0,105612	0,660000
0,280340	0,980392	0,105125	0,673333
0,279095	0,980769	0,104653	0,686667

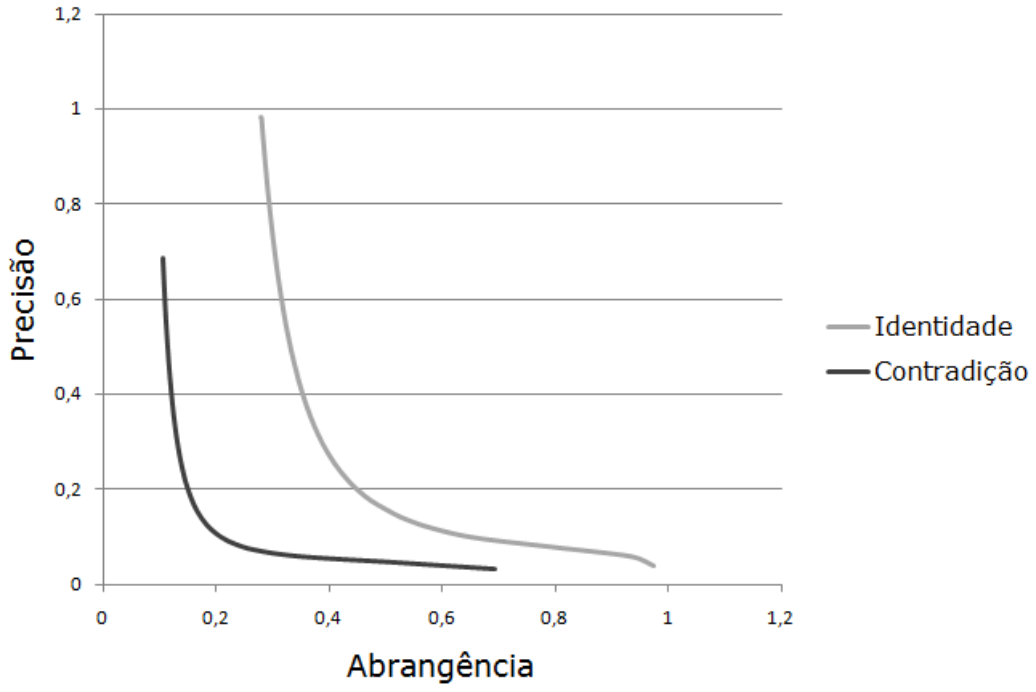

Figura 14. Resultados da avaliação do protótipo do modelo Vetorial estendido

5.1.5 Análise dos resultados

Nota-se maior precisão nas relações do tipo identidade. Um motivo para essa característica é que a definição é próxima do que originalmente fora estabelecido no modelo CST e esse tipo de correlação é a de menor subjetividade (AFANTENOS, DOURA, *et al.*, 2004).

A pouca identificação de contradições pode estar relacionada à possível limitação do arcabouço linguístico quanto aos antônimos. Outro questionamento é que ainda não são tratadas questões dos predicados na forma de numerais, datas e outros atributos valorados, por exemplo, “O avião caiu e trinta pessoas morreram” e “O avião caiu e somente vinte pessoas faleceram”.

5.2 Fique sabendo: um sistema de disseminação seletiva da informação

O objetivo desta seção é explicitar e elaborar o protótipo de um sistema de disseminação seletiva da informação que atente tanto para os requisitos enunciados, Capítulo 2, quanto para a seleção de conteúdo por meio do modelo Vetorial proposto, Capítulo 3 e 4.

A intenção não é construir um sistema de máxima completude, mas que permita constatar as proposições anteriores, enquanto protótipo computacional. Desse modo, alguns dos critérios não foram explorados por completo. No entanto, perspectivas de melhorias foram expostas para a continuidade e para o desenvolvimento do sistema em futuros trabalhos (ver Seção 6.2).

Na próxima seção são relatados os requisitos do sistema, que se basearam nas condições inicialmente exploradas no Capítulo 2 e revisitadas de maneira a ressaltar a importância desses conceitos para sistemas de disseminação seletiva da informação.

Posteriormente, são apresentadas a visão geral do sistema, as funções básicas construídas e as interfaces. Por último, são feitas avaliações do sistema como aquelas realizadas para outros sistemas no Capítulo 2.

5.2.1 Especificação dos requisitos

Os requisitos elicitados para a concepção do protótipo representam os requisitos funcionais e não funcionais do sistema de informação, optando por não se fazer distinção desses, de modo a privilegiar pela organização das ideias, segundo o exposto no Capítulo 2.

Quanto à forma de obtenção, o sistema deverá:

i. Fornecer descritores interoperáveis, por exemplo, por meio de linguagem de marcação de dados como aquelas fornecidas pela Web Syndication;

ii. Extrair informação de conteúdos textuais, para compor as triplas proposicionais do grafo conceitual;

iii. Gerar descritores que contemplem as correlações semânticas entre discursos, no processo de indexação expandida e pesquisa vetorial.

Quanto ao fomento do conteúdo, o sistema suportará os tipos:

i. Estrutural: para guardar dados sobre as correlações dos documentos, por exemplo, tipo, sentenças, documentos;

ii. Vetorial: de acordo com o modelo proposto, o sistema se comporta como campo vetorial capaz de identificar correlações semânticas básicas;

iii. Rede: não será observado por esta dissertação, porém foi dito nas seções anteriores que o modelo de rede poderá ser concernido à solução, à medida que as correlações semânticas entre documentos forem sendo instanciadas, formando uma rede semântica de documentos;

iv. O mapeamento semântico em nível de entidades ainda não será abordado, mas a tendência é que os meta-dados também sejam relacionados em trabalhos futuros, visto que os atributos dos meta-modelos também podem assumir valores de textos livres.

Quanto ao tipo de interesse, o sistema se equiparará:

i. Explícito: considera-se que o próprio discurso dos agentes na interface de interação propiciará a explicitação do interesse, correlacionando o documento às produções de outros agentes;

ii. Implícita: o sistema irá inferir outros termos prováveis de interesse do agente (usuário), por meio do arcabouço linguístico e dos grafos conceituais.

A abordagem será complexa, à medida que o interesse não será estático. O tipo de seleção será parcial, obtendo os resultados similares ao interesse, de acordo com cada correlação. Porém, dado que os documentos são relacionados semanticamente, a seleção também será relacional.

Como serão selecionados documentos de diferentes agentes (usuário), podem-se propor meios para complementar o sistema de modo a suportar mecanismos de seleção social, por exemplo, indicações.

Quanto à qualidade, o sistema deverá permitir retroalimentação e gerência do interesse. A veracidade da informação não será tratada nessa versão inicial, mas a possibilidade de exploração estatística das relações poderá ser realizada em trabalhos futuros.

O sistema automaticamente fará a síntese dos resultados, uma vez que exibirá somente as sentenças que se correlacionam e não todo o texto do discurso. Não haverá módulos específicos para interface com especialistas em mediação (SOUTO, 2008) (SOUTO, 2008), porém se espera que essa seja intuitiva ao ambiente, pela facilidade de interação do sistema e dos artefatos que se apresentarão.

Quanto à interoperabilidade, o sistema deverá prover canais de comunicação com outros sistemas, por meios das tecnologias atuais de integração. Essa característica é essencial para a escalabilidade da ferramenta em diferentes ambientes.

Pelo fato de se trabalhar com discursos textuais, a ferramenta terá uma curva de aprendizagem baixa, reduzindo o esforço cognitivo. A facilidade de interação é inerente a essa perspectiva.

Ainda que o objetivo da ferramenta não seja específico para apoio à aprendizagem, não deverá essa se redimir da responsabilidade de tratar o acesso a informação como um dos caminhos para a construção de conhecimento. Desse modo, a intenção ao se utilizar (FREIRE, 1999) como referencial teórico é promulgar reflexões sobre como os sistemas de disseminação seletiva são concebidos, ao ignorarem a atividade construtiva.

Diante dos vários princípios abordados na referência, esta dissertação se limitará a mostrar que, de um cenário em que há receptores de informação, inaptos a exposição da criticidade, pode-se transformar um sistema de disseminação para que ele contemple a prática do discurso e da interação com o ambiente social provido, o que facilitara a aprendizagem de novos conhecimentos, promovendo outros horizontes para esse tipo de sistema.

5.2.2 Visão geral do sistema

O sistema é composto por uma interface de captação de conteúdo, em que os agentes submetem textos que refletem algo que se queira disseminar, por exemplo, uma opinião, um fato ou algo novo produzido por outras fontes de informação.

O texto produzido será base para a composição do interesse do agente (usuário), o que posteriormente fará com que sejam selecionados outros textos, de diferentes autores, em que as correlações semânticas foram estabelecidas pelo sistema.

O módulo fornece também serviços de comunicação para que essas informações possam ser enviadas por outros sistemas. As interfaces estão ilustradas nas Figuras 15 e 16. As tecnologias de comunicação utilizadas são aquelas consolidadas em sistemas distribuídos, sendo utilizada a biblioteca Apache Camel[5].

[5] http://camel.apache.org/

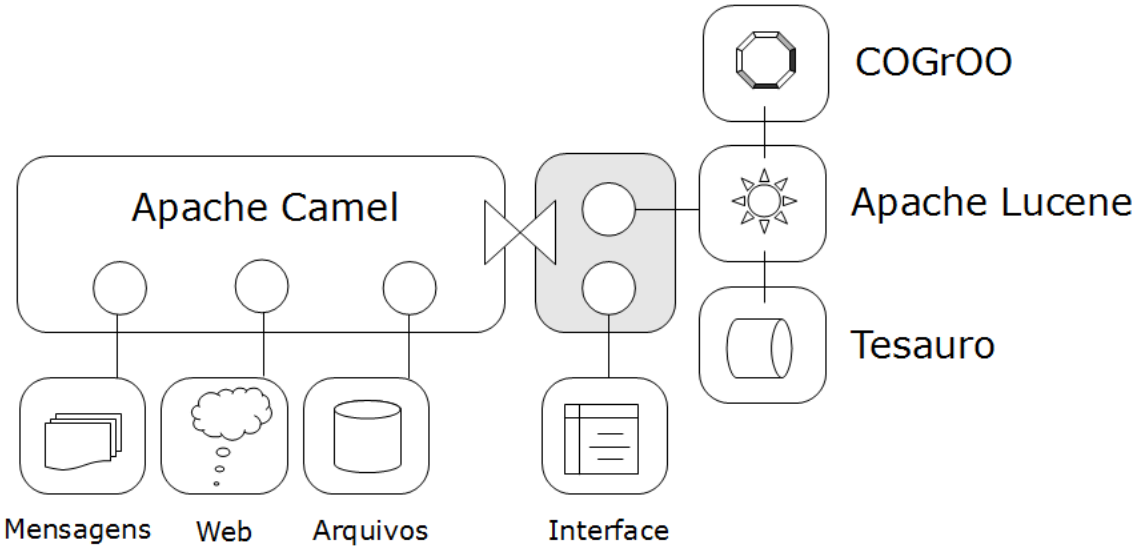

Figura 15. Arquitetura geral do sistema

Figura 16. Interface simples de captação de conteúdo

Os novos conteúdos são processados pelo módulo Vetorial, em que são identificadas as correlações semânticas e são indexadas as triplas conceituais. Nesse momento, o interesse é mapeado da seguinte forma: o sujeito de maior relevância (que possui o maior número de filhos e menor número de pais) e o predicado de maior relevância (regra inversa do sujeito) serão candidatos a termos de interesse, sendo consultados no arcabouço linguístico as identidades correspondentes.

O conjunto de termos formado pela junção desses compõe a lista de interesse, sendo essa gerenciada pelo agente em outra interface. Podem-se estabelecer critérios de ordenamento dessa lista, por exemplo, termos com mais tempo sem serem mencionados perdem valor (Figura 17). Tanto a pesquisa da tripla conceitual, quanto à pesquisa dos termos da lista de interesse geraram outra lista dinâmica contendo os documentos correlacionados.

Figura 17. Representação e gerência do interesse

O conjunto de documentos forma uma rede de ligações para outros documentos, em que é possível se trafegar entre os diferentes conteúdos. Cada rede de correlação semântica pode ser instanciada na forma de canais de interesse, sendo possível instanciar os resultados em arquivo no formato de marcação da *Web Syndication* (GOLBECK e HALASCHEK-WIENER, 2009). Assim, se o acesso ao arquivo é público a rede de computadores, então os conteúdos são disseminados seletivamente. Esses conceitos são ilustrados na Figura 18.

Figura 18. Interface geral do sistema

5.2.3 Avaliação do sistema

A avaliação do sistema construído, assim como fora realizado para os quatro sistemas relevantes no Capítulo 2, será subjetiva e visa evidenciar os conceitos discutidos por este trabalho. Ao final desta avaliação, o sistema também foi pontuado (Tabela 8 – número *v*). Para melhor entendimento dos critérios utilizados, separou-se a avaliação em quatros elementos a se destacar (Figura 19):

Figura 19. Elementos a se destacar no protótipo

1. O sistema possui uma interface intuitiva e de fácil interação. Isso reduz a curva de aprendizagem (esforço cognitivo) da ferramenta. Uma evolução prevista do sistema (ver Seção 6.2) será trabalhar também com hiper-textos ou hiper-mídias, desde que seja possível obter o conteúdo em formato textual;

2. A representação do interesse é implícita, porém, exibida para conferência e gestão, logo também explícita. Dessa forma, são evidentes, para quem se utiliza do sistema, quais foram os critérios de seleção de conteúdo. No protótipo, fora utilizada a técnica *Tag Cloud*, descrita em (KUO, HENTRICH, *et al.*, 2007);

3. O ícone da figura é um atalho para um arquivo com formato de marcação de meta-dados do tipo *Web Syndication*. Logo, o resultado da seleção de conteúdo poderá ser disseminado para outros sistemas, como os descritos em (ALMEIDA, 2008);

4. Outros documentos que possuem correlação semântica também estão acessíveis e são exibidos os contatos dos autores. Tal característica incentiva a interação entre os participantes do sistema. Uma evolução prevista do protótipo (ver Seção 6.2) é a possibilidade de criação de operadores de seleção (*templates*) como os que foram elaborados em (JORGE, 2010).

Ao final da construção do protótipo, ressalta-se (Capítulo 2): devem os sistemas de disseminação seletiva observar abordagens centradas nos interesses e nas necessidades, não somente de dados e de informações, mas em novas possibilidades de socialização. Um campo de pesquisa promissor para isso é a análise do discurso, voltado à obtenção de correlações intertextuais. O mérito desta dissertação foi exemplificar que essa perspectiva é tangível.

Tabela 8. Síntese dos resultados de avaliação do protótipo

	Critério	**i**	**ii**	**iii**	**iv**	**v**
1	Quando os descritores são fornecidos	0	2	1	2	1
2	Os índices são extraídos do conteúdo	2	0	0	0	2
3	Os descritores são inferidos	1	0	0	0	2
4	Utiliza-se de bancos de dados	0	0	2	0	1
5	Utiliza-se de modelos vetoriais	2	0	0	0	2
6	Utiliza-se de modelos de rede	1	2	0	1	1
7	Utiliza-se de modelos de meta-dados	0	2	0	2	1
8	Interesse dirigido pelo conteúdo	1	2	0	2	1
9	Interesse é expresso pelo receptor	1	1	2	1	2
10	Interesse é inferido pelo sistema	2	0	0	0	2
11	Não utiliza processos evolutivos	0	2	2	2	1
12	Utiliza evolução ou expansão do interesse	1	0	0	0	2
13	Os resultados correspondem à pesquisa	0	2	2	2	1
14	Os resultados são similares à pesquisa	2	0	0	0	2
15	Seleção devido à relação entre documentos	1	2	0	2	1
16	Seleção devido a critérios sociais do receptor	1	0	1	0	1
17	O sistema avalia resultados (retroalimentação)	1	0	0	0	2
18	O sistema valida índices e documentos	1	0	0	0	1
19	O sistema sintetiza os resultados	1	2	0	2	2
20	O sistema possui interface para mediação humana	1	0	2	0	1
21	O sistema integra facilmente com outros sistemas	1	2	0	2	2
22	A curva de aprendizagem do sistema é baixa	1	1	1	1	2
23	A interação é mediada e intuitiva	1	0	2	1	2
24	O sistema auxilia na aprendizagem construtiva	0	0	0	0	1
	Total	**22**	**20**	**15**	**20**	**36**

5.3 Outras aplicações

As aplicações de disseminação seletiva estão presentes também em outros sistemas de informação, construídas na forma de componentes de *software* acopláveis. Ao ponto que atendem ao requisito de propagação ou de acesso personalizado a conteúdos produzidos nesses ambientes, o objetivo dessa seção é investigar como o sistema de disseminação proposto por este trabalho poderá oferecer, além desses serviços, novos meios facilitadores dos processos de interação e de cooperação.

5.3.1 Aplicações em sistemas colaborativos

Dentre as classes de sistemas de informação atuais, esta seção buscou analisar como os conceitos trabalhados por esta dissertação podem ser usados em ambientes específicos para realização de atividades colaborativas.

Novos estudos sobre sistemas colaborativos estão sendo propostos pela comunidade científica, principalmente na elaboração de modelos para sistematização das atividades (OLIVEIRA, 2009).

Independentemente de como os sistemas são modelados, pode-se afirmar que mecanismos de interação e de cooperação são essenciais nesses ambientes. A comunicação entre os participantes é uma das etapas para realização dessas atividades e, nesse ponto, os sistemas de disseminação seletiva agregam meios para propagação e para seleção de conteúdo.

No entanto, podem-se eleger outros assuntos em que o sistema de disseminação seletiva proposto por esta dissertação possa ser útil. Portanto, um mérito deste capítulo é relatar aplicações do modelo proposto. Desse modo, enumeram-se a seguir algumas indicações:

i. Aproximação social e estímulo a interação: nem sempre todos os participantes dos sistemas colaborativos se conhecem ou se interagem (PRIMO, 2007). Assim, um dos desafios dos sistemas é criar condições para que os participantes com objetivos ou interesses afins possam se relacionar, ao que se designa de aproximação social. Alguns sistemas desenvolveram funcionalidades de comparação entre perfis para identificar afinidades e fazer sugestões de contato (CHEN, GEYER, *et al.*, 2009). A

identificação de correlações entre os conteúdos pode ser uma alternativa para que essa propriedade se torne constante, à medida que é ofertado ao participante que outras pessoas também produziram documentos relacionados ao conteúdo dele. Por exemplo, quando duas ou mais pessoas comentam sobre um mesmo assunto em páginas pessoais, se essas informações forem confrontadas automaticamente pelo ambiente, não haveria a necessidade dos autores procurarem se o conteúdo já fora debatido ou se terá alguém interessado em lê-lo ou até em respondê-lo. Para tanto, a exibição de que o novo conteúdo publicado possui correlação semântica com documentos de outras pessoas poderia induzir o autor a um debate de opiniões com as pessoas que manifestadamente também tem interesse naquele tema (Figura 20).

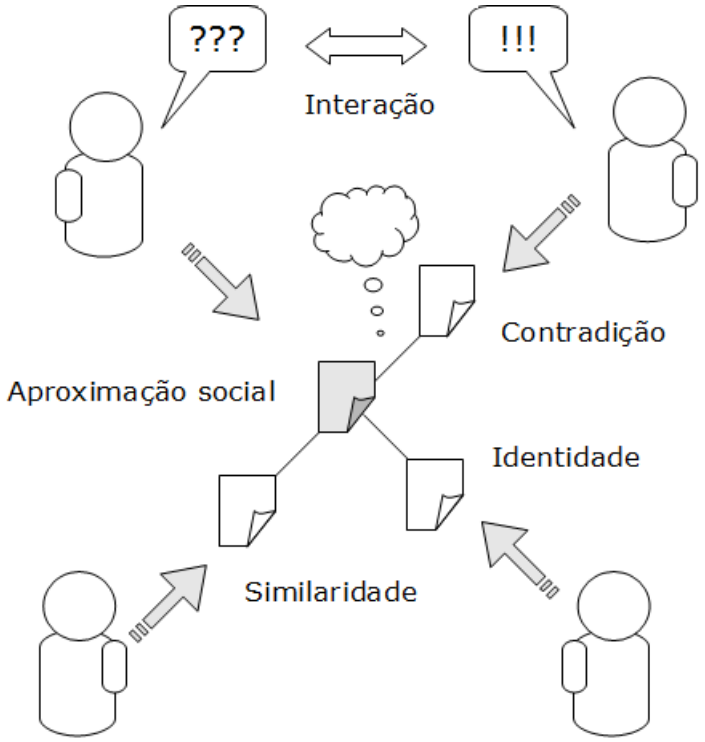

Figura 20. Aproximação social e estímulo a interação

ii. Identificação de comunidade de especialistas: em sistemas colaborativos, é comum a necessidade de se identificar especialistas em assuntos. Dentre as propriedades providas pela correlação semântica de documentos, nota-se que, se um agente ou comunidade de agentes (usuário) possuírem documentos muito correlacionados, então é alta a probabilidade de se tratar de especialista no tema;

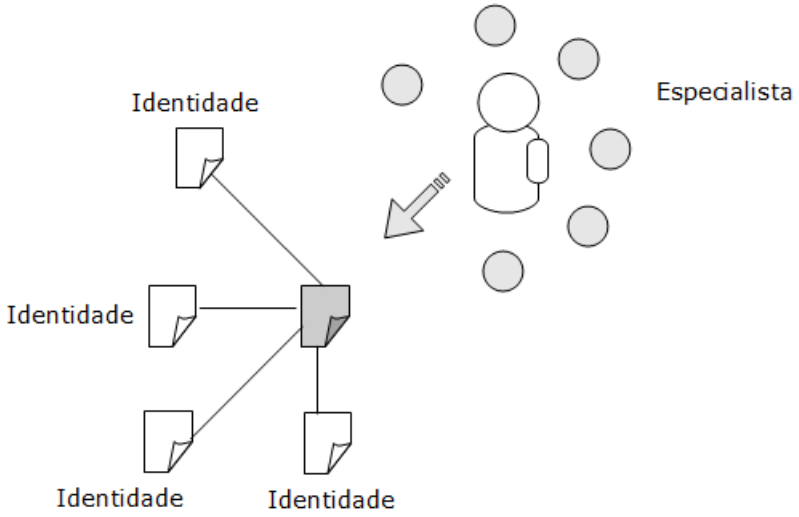

Figura 21. Identificação de comunidade de especialistas

iii. Melhoria de percepção das atividades em grupo: do que é relatado nos itens, nota-se que os sistemas de disseminação seletiva auxiliam na percepção de "qual é" e de "como está" o trabalho do grupo em que o agente está envolvido. Seja transmitindo seletivamente o conteúdo das produções coletivas, seja correlacionando o conteúdo do participante no sistema colaborativo, a disseminação seletiva tem um papel que é além de um propulsor de informações.

5.3.2 Aplicações em informática na educação

De maneira semelhante à última seção, procurou-se investigar como a abordagem do modelo de disseminação proposto por este trabalho é conduzida a ambientes de ensino e aprendizagem.

A potencial apropriação do modelo em sistemas colaborativos corrobora com a ideia de que sistemas de disseminação seletiva podem ser usados para interações além das simples atividades de comunicação.

Nesse sentido, pode-se direcionar também essa pesquisa à proposição de novas formas de cooperação, por meio do algoritmo proposto por este trabalho. A fim de exemplificar como esta solução tem aplicabilidade em informática na educação, limitou-se o trabalho a duas linhas de pesquisa, uma em tutores inteligentes, outra em arquiteturas pedagógicas:

i. Tutores inteligentes sociointeracionistas: tutores inteligentes é um tipo de sistema para auxílio à educação que modela propostas pedagógicas aliadas a domínios de conhecimento para inferir sobre o modo de compreensão do aluno, adaptando individualmente o ensino as necessidades (VANLEHN, 1988). De acordo com a concepção clássica de tutores inteligentes, considera-se somente a interação de um aluno por máquina, o que é um fator limitante na apropriação desses ambientes. A dificuldade de se ter cenários para múltiplos aprendizes é o quanto complexo se faz a tarefa de modelar interesses e necessidades coletivas. Portanto, pode-se utilizar dos conceitos de correlação semântica entre documentos produzidos pelos alunos para relacionar interesses em comum, não mais individualizados, e propor atividades de forma colaborativa;

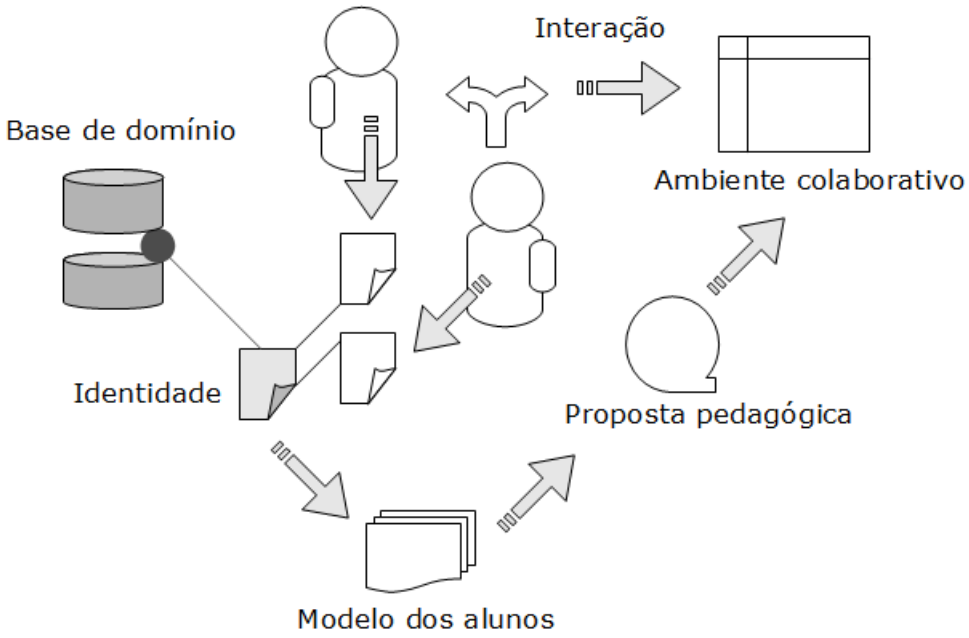

Figura 22. Tutores inteligentes sociointeracionistas

ii. Mediação em "Controvérsia Acadêmica" (JOHNSON e JOHNSON, 1994): da concepção de arquiteturas pedagógicas (CARVALHO, NEVADO e MENEZES, 2005), insere-se a "Controvérsia Acadêmica" como metodologia de aprendizagem que incentiva interações e debates quando produções intelectuais dos aprendizes são incompatíveis. Portanto, o modelo de correlações semânticas pode ser usado como mediação tecnológica: ao identificar contradições entre discursos, acionam-se interfaces para discussões entre os alunos e para acompanhamento da aprendizagem

pelo professor. Nesse cenário, os sistemas podem ser vistos como componentes ou como bibliotecas a serem utilizados por ambientes flexíveis, dentre eles, o MOrFEu (MENEZES, NEVADO, *et al.*, 2008), (BELTRAME, CURY, *et al.*, 2008), (RANGEL, BELTRAME, *et al.*, 2009), (SANTOS, CASTRO e MENEZES, 2010). Um dos objetivos desses ambientes é o suporte telemático a diferentes arquiteturas pedagógicas (RANGEL, 2011). Apesar de esta dissertação exemplificar a aplicação do modelo de correlações semânticas entre discursos somente para a arquitetura "Controvérsia Acadêmica", é recomendado que futuros trabalhos investiguem e proponham novas abordagens da metodologia (ver Seção 6.2) nesses ambientes de aprendizagem.

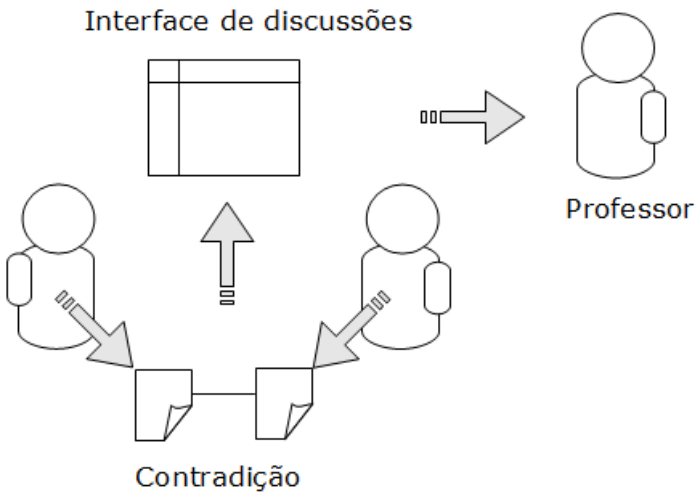

Figura 23. Mediação em "Controvérsia Acadêmica"

5.4 Considerações parciais

Resumidamente, os enfoques deste capítulo foram: apresentar detalhes do estudo de casos realizado após a concepção do modelo Vetorial do capítulo anterior; evidenciar a aplicação da proposta em sistemas de disseminação seletiva, com a premissa de abordar os conceitos avaliados no Capítulo 2; exemplificar outras possíveis aplicações do modelo em dois tipos de sistemas de informação, os sistemas colaborativos e os ambientes de aprendizagem.

Os resultados experimentais do modelo Vetorial podem ser considerados satisfatórios, com a ressalva de que uma melhor eficácia dependerá da evolução do arcabouço linguistico e da melhoria do módulo gerador de grafo conceitual.

Com a prerrogativa de se elaborar um protótipo, nem todos os requisitos de um sistema de um sistema de disseminação seletiva da informação (Capítulo 2) foram desenvolvidos por completo. No entanto, diretrizes para que novos trabalhos aprimorem a ferramenta foram expostas e discutidas.

Os exemplos que foram relatados na última seção carecem de aprofundamento, tanto teórico, quanto de experimentações. Todavia, espera-se que essa atividade seja realizada pela continuidade desta proposta (ver Seção 6.2).

6 CONSIDERAÇÕES FINAIS

Este trabalho teve como mérito principal a proposição da "análise de múltiplos discursos" com finalidade de compor um sistema de disseminação seletiva da informação. A análise de múltiplos textos foi baseada na teoria *Cross-Document Structure Theory*. Utilizou-se de técnicas de geração automática de grafos conceituais a partir de textos, para elaborar um modelo Vetorial estendido que fosse capaz de identificar correlações semânticas básicas.

Os atuais sistemas de disseminação seletiva apresentam deficiências quanto à estruturação dos interesses dos usuários. Foi proposto que, com a adoção do modelo, tais sistemas podem ser aperfeiçoados de modo que contemplem às necessidades, não somente de dados e de informações, mas de novas possibilidades de socialização.

6.1 Objetivos alcançados

Considera-se que um dos objetivos alcançados por esta dissertação diz respeito à proposta de evolução dos recentes sistemas de disseminação seletiva da informação para cenários de seleção de conteúdo baseada em correlações entre discursos.

Mostrou-se que tal iniciativa é passível de desenvolvimento, à medida que foi apresentado um protótipo computacional, com resultados satisfatórios a avaliação definida segundo os critérios do próprio trabalho, ainda que nem todos os requisitos tenham sido atendidos.

Destaca-se como contribuição deste trabalho a apresentação de nova solução vetorial quanto à recente teoria *Cross-Document Structure Theory* e outra utilização do método para além do uso em sumarização de documentos, voltando-se a característica potencial de seleção de conteúdo nos sistemas de disseminação.

Diferentemente das avaliações feitas originalmente nas referências sobre o assunto, preferiu-se utilizar como métricas os critérios Precisão e Abrangência, comum em sistema de recuperação da informação. Após avaliação em base anotada manualmente, ainda que com número não grande de documentos, o escopo inicial de verificar a aplicabilidade da solução foi atingindo. Sendo expostos os resultados, espera-se que novas propostas os utilizem como parâmetros quanto à abrangência, à complexidade e ao desempenho.

6.2 Trabalhos futuros

Em várias partes do texto é mencionada a possibilidade de evolução do que fora exposto. Cabe, então, reapresentar as passagens e propor sugestões de continuidade do trabalho:

i. Cabe a trabalhos futuros relatar as diferentes formas de se compor um arcabouço linguistico e qual das áreas de pesquisa, europeia ou norte-americana, oferecerão apoio para as diversas necessidades e cenários que possam existir. Além dessas necessidades, um trabalho futuro se faz necessário para formalizar a definição do termo “arcabouço linguistico” e investigar trabalhos correlatos;

ii. Um estudo mais aprofundado sobre os pesos dos vetores é necessário. Da mesma forma que no modelo Vetorial clássico, as funções de ponderação são essenciais para encontrar melhores resultados. Variações dos pesos devem ser avaliadas por meio de testes e de experimentações;

iii. A proposta utilizou, ainda que intuitivamente, de conceitos da lógica proposicional. Porém, ao referir-se como solução vetorial, atentou-se que a junção dessas visões poderá ser trabalho de aprofundamento teórico. Portanto, um possível trabalho futuro deverá fazer levantamentos das correspondências entre as abordagens e propor melhorias da solução;

iv. A prospecção para modelos associativos em rede é um dos pontos a serem trabalhados e novas propostas poderão investigar o uso estatístico da rede semântica para validar e invalidar correlações, ainda que a rede semântica seja construída de forma automática. Deverão ser investigadas técnicas de topologia de grafos para extrair informações entre conceitos, documentos e sociabilidade;

v. Para validar os grafos gerados foi elaborada uma interface gráfica capaz de exibir e manipular grafos. Porém, a interface e os componentes utilizados não foram detalhados. Deverá um trabalho futuro especificar esses elementos e avaliar a utilização em outros cenários;

vi. No protótipo só foram tratadas as anáforas pronominais. Outros tipos de anáforas podem ser trabalhados. É necessário ainda verificar a proposta de uso dos grafos conceituais para determinação dos elementos centrais de um texto;

vii. Para validação do arcabouço linguistico, também foi construída uma interface gráfica. Porém, a interface e os componentes utilizados não foram detalhados. Deverá um trabalho futuro especificar esses elementos e avaliar a utilização em outros cenários;

viii. Para validação das correlações semânticas, também foi construída uma interface gráfica. Porém, a interface e os componentes utilizados não foram detalhados. Deverá um trabalho futuro especificar esses elementos e avaliar a utilização em outros cenários;

ix. Uma evolução do protótipo do sistema de disseminação seletiva da informação será trabalhar também com hiper-textos ou hiper-mídias, desde que seja possível obter o conteúdo em formato textual. Ferramentas e algoritmos de conversão deverão ser avaliados;

x. Outra evolução do protótipo do sistema de disseminação seletiva da informação será possibilidade de criação de operadores de seleção (*templates*). É necessário investigar se os operadores poderão ser usados também para seleção de conteúdo voltada à disseminação seletiva;

xi. Ainda que o modelo de correlações semânticas entre textos fora exemplificado para a arquitetura pedagógica “Controvérsia Acadêmica”, recomenda-se que futuros trabalhos investiguem novas abordagens da metodologia em ambientes flexíveis que suportem a composição de outras arquiteturas.

6.3 Conclusões

Buscou-se, durante a escrita deste trabalho, pela organização das ideias, de forma resumida e objetiva. Ainda que caracterizada como exploratória e descritiva, esta dissertação teve um caráter mais descritivo, eximindo-se de revisões literárias à exaustão.

Assume-se que as hipóteses foram convalidadas e os objetivos iniciais alcançados. Porém, pode-se concluir que o modelo apresentado possui necessidades de aperfeiçoamento e de fundamentação.

Uma parte dessas necessidades está explicitada na seção anterior (trabalhos futuros) e a outra intrínseca nas observações relatadas no decorrer da dissertação. Espera-se que essas diligências sejam motivadoras para a continuidade de revisões e de novas propostas.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AFANTENOS, S. D. et al. Exploiting Cross-Document Relations for Multi-document Evolving Summarization. **SETN**, 2004. 410-419.

AGOSTI, M.; MARCHETTI, P. G. User navigation in the IRS conceptual structure through a semantic association function. **RIK**, 1992.

ALEIXO, P.; PARDO, T. A. S. Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. **Workshop em Tecnologia da Informação e da Linguagem Humana – TIL**, 2008. 298-303.

ALMEIDA, R. L. D. **Disseminação de Conteúdo na Web:** A tecnologia RSS como Proposta para Comunicação Científica. Brasília: Universidade de Brasília, 2008. Dissertação de Mestrado.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1. ed. New York: ACM Press books, 1999.

BAX, M. P. et al. Sistema Automático De Disseminação Seletiva. **IFLA**, São Paulo, 2004.

BECKER, J. Topic-based Vector Space Model. **BIS**, Colorado, 2003.

BEIGBEDER, M. Integrating Boolean and vector models of information retrieval with passage retrieval. **WISICT**, Dublin, 2005.

BELTRAME, W. A. R. et al. Multi-Organizador Flexível de Espaços Virtuais. **SBIE**, 2008.

BROWN, P. F.; LAI, J. C.; MERCER, R. L. Aligning sentences in parallel corpora. **ACL**, 1991.

CARVALHO, M. J. S.; NEVADO, R. A.; MENEZES, C. S. Arquiteturas Pedagógicas para Educação a Distância: Concepções e Suporte Telemático. **SBIE**, 2005.

CHEN, J. et al. Make new friends, but keep the old: recommending people on social networking sites. **Conference on Human Factors in Computing Systems**, Boston, 2009. 201-210.

DEEWESTER, S. et al. Indexing by Latent Semantic Analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391-407, 1990.

DIAS-DA-SILVA, B. C.; FELIPPO, A. D.; NUNES, M. D. G. V. The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database. **LREC**, 2008.

DIESTEL, R. **Graph Theory**. New York, USA: Eletronic, 2005.

EIRÃO, T. G. Disseminação Seletiva da Informação: Uma Abordagem. **Revista Digital de Biblioteconomia e Ciência da Informação**, 2009.

EIRÃO, T. G. **A disseminação seletiva da informação e a tecnologia RSS nas bibliotecas de Tribunais em Brasília**. Brasília: Universidade de Brasília, 2011.

FERREIRA, J.; SILVA, A. **MySDI:** A Generic Architecture to Develop SDI Personalised Services (How to Deliver the Right Information to the Right User?). Setubal: ICEIS, 2001.

FREIRE, P. **Pedagogia da autonomia:** saberes necessários à prática educativa. São Paulo: Paz e Terra, 1999.

FREITAS, S. A. A. D. **Interpretação Automatizada de Textos:** Processamento de Anafóras. Vitória: Universidade Federal do Espírito Santo, 2005. Tese de Doutorado.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, 1997. ISSN 55(1):119-139.

GOLBECK, J.; HALASCHEK-WIENER, C. Trust-based Revision for Expressive Web Syndication. **Journal of Logic and Computation**, v. 19, n. 5, p. 771-790, 2009.

GUIZZARDI, G. **Ontological Foundations for Structural Conceptual Models**. Twente: University of Twente, 2005.

HASSAN, A. et al. **Content based recommendation and summarization in the blogosphere**. California: ICWSM, 2009.

HEARST, M. A. et al. Support vector machines. **Intelligent Systems and their Applications, IEEE**, v. 13, n. 4, 1998. ISSN 1094-7167.

JBARA, A. A.; RADEV, D. R. Coherent citation-based summarization of scientific papers. **HLT**, 2011.

JOHNSON, D. W.; JOHNSON, R. Structuring Academic Controversy. **Handbook of cooperative learning methods**, 1994.

JORGE, M. L. D. R. C. **Sumarização automática multidocumento:** seleção de conteúdo com base no modelo CST(Cross-Document Structure Theory). São Paulo: Universidade de São Paulo, 2010. Dissertação de Mestrado.

KANSA, E. C.; BISSELL, A. Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains. **Data Science Journal**, 2010.

KAWAHARA, D.; INUI, K.; KUROHASHI, S. Identifying contradictory and contrastive relations between statements to outline web information on a given topic. **COLING**, Stroudsburg, 2010.

KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. CoGrOO - An OpenOffice Grammar Checker. **Seventh International Conference on intelligent Systems Design and Applications - ISDA**, Washington, 2007.

KOWATA, J. H. **Uma Abordagem Computacional para a Construção de Mapas Conceituais a partir de Textos em Língua Portuguesa do Brasil**. Vitória: Universidade Federal do Espírito Santo, 2010. Dissertação de Mestrado.

KRISHNA, M. et al. **The effect of linguistic constraints on the large scale organization of language**. Arxiv. New York. 2011.

KUO, B. Y.-L. et al. Tag clouds for summarizing web search results. **Conference on World Wide Web**, Banff - Canadá, 2007. 1203 - 1204.

LANCASTER, F. W. **Indexação e Resumos:** Teoria e Prática. São Paulo: Briquet de Lemos, 2004.

LAURENCE, H.; HIRSCH, R.; SAEEDI, M. Evolving Lucene search queries for text classification. **GECCO**, 2007.

LAZARTE, L. Ecologia cognitiva na sociedade da informação, Brasília, 2000.

LUHN, H. P. Selective dissemination of new scientific information with the AID of electronic processing equipment. New York: American Documentation, 1961.

MANN, W. C.; THOMPSON, S. A. **Rhetorical Structure Theory: A Theory of Text Organization**. ISI Reprint Series. New York. 1987. (ISI/RS-87-190).

MEDEIROS, J. D. S. **Tesauros conceituais e ontologias de fundamentação:** modelos conceituais para representação de domínio. Rio de Janeiro: Universidade Federal Fluminense, 2011. Dissertação de Mestrado.

MENEZES, C. S. et al. MOrFEu – Multi-Organizador Flexível de Espaços Virtuais para Apoiar a Inovação Pedagógica em EAD. **SBIE**, 2008.

MILLER, G. A. WordNet: An Electronic Lexical Database. **Communications of the ACM**, v. 38, n. 11, p. 39-41, 1995.

MONTEIRO, E. R. **RSN - Rede Social de Notícias**. Vitória: Universidade Federal do Espírito Santo, 2009. Dissertação de Mestrado.

MORALES-DEL-CASTILLO, J. M. et al. **A Semantic Model of Selective Dissemination of Information for Digital Libraries**. New York: Information Technology and Libraries, 2009.

MURAKAMI, K. et al. Statement map: assisting information crediblity analysis by visualizing arguments. **WICOW**, Madri, 2009. 43-50.

MURAKAMI, K. et al. Automatic Classification of Semantic Relations between Facts and Opinions. **NLPIX**, 2010.

OLIVEIRA, F. F. D. **Uma Ontologia de Colaboração e suas Aplicações**. Vitória: Universidade Federal do Espírito Santo, 2009. Dissertação de Mestrado.

PAPAEMMANOUIL, O.; ÇETINTEMEL, U. SemCast: Semantic Multicast for Content-based Data Dissemination. **ICDE**, 2004.

PARSAYE, K. et al. **Intelligent databases:** object-oriented, deductive hypermedia technologies. New York: John Wiley & Sons, 1989.

PETROVIC, M.; LIU, H.; JACOBSEN, H.-A. G-ToPSS: Fast Filtering of Graph-based Metadata. **IW3C2**, 2005.

PRIMO, A. **Interação mediada por computador:** comunicação, cibercultura, cognição. São Paulo: Sulina, 2007.

QAZVINIAN, V.; RADEV, D. R. **Identifying non-explicit citing sentences for citation-based summarization**. New York: ACL, 2010.

RADEV, D. R. A common theory of information fusion from multiple text sources step one: cross-document structure. **WDI**, 2000.

RANGEL, V. G. **VCom:** uma Abordagem para a Modelagem de Ambientes Colaborativos. Vitória: Universidade Federal do Espírito Santo, 2011. Dissertação de Mestrado.

RANGEL, V. G. et al. MOrFEu:Towards the Design of an Environment for Flexible Virtual Spaces Organization. **WCCE – World Conference on Computer in Education**, 2009.

SALTON, G.; WONG, A.; YANG, C. S. Vector Space Model for Automatic Indexing. **ACM**, New York, v. 18, n. 11, 1975.

SANTOS, L. N.; CASTRO, A. N. J.; MENEZES, C. S. MOrFEu: Criando Ambientes Virtuais Flexíveis na Web para mediar a Colaboração. **IE - Congreso Iberoamericano de Informática Educativa**, 2010.

SARACEVIC, T. Ciência da Informação: Origem, Evolução e Relações. **Perspectiva em Ciência da Informação**, v. 1, n. 1, 1996.

SOUTO, L. F. **Mediação em serviços de disseminação seletiva de informações no ambiente de bibliotecas digitais federadas**. São Paulo: Universidade de São Paulo, 2008. Tese de Doutorado.

SOUZA, F. S. L. D. et al. Uma Abordagem para Comparação de Mapas Conceituais utilizando Correspondência de Grafos. **RENOTE - Revista Novas Tecnologias na Educação**, v. 4, n. 2, 2006. ISSN 1679-1916.

SPARCK-JONES, K. Assumptions and issues in text-based retrieval. **JAC**, 1992.

TRIGG, R. H. **A Network-based Approach to Text Handling for the On-line Scientific Community**. Maryland: University of Maryland, 1983. Ph. D. thesis.

TRIGG, R. H.; WEISER, M. **TEXTNET:** A Network-based Approach to Text Handling. New York: ACM Transactions on Information Systems (TOIS), 1986.

VANLEHN, K. Student modeling. In: POLSON, M. C.; RICHARDSON, J. J. **Foundations of intelligent tutoring systems**. [S.l.]: Routledge, 1988. p. 55 -78.

ZHANG, Z.; OTTERBACHER, J.; RADEV, D. Learning Cross-document Structural Relationships using Boosting. **CIKM**, Louisiana, USA, 2003.

ZHANG, Z.; RADEV, D. R. Learning cross-document structural relationships using both labeled and unlabeled data. **IJC-NLP**, Hainan Island, China, 2004.