Texto & Contexto: por uma recuperação da informação com mais semântica

Ulrich Schiel
Pós-Doutorado pela Gesellschaft Für Mathematik Und Datenverarbeitung (GMD) - Alemanha. Doutor em Informática pela Universität Stuttgart (UNI-STUTTGART) - Alemanha. Professor da Universidade Federal da Paraíba (UFPB) - Brasil.
http://lattes.cnpq.br/2971250918247087
E-mail: ulrich@computacao.ufcg.edu.br

Data de submissão: 01/02/2021. Data aceite: 17/09/2021. Data de publicação: 31/12/2021

RESUMO

Com o advento da World Wide Web, a Recuperação da Informação cresceu em significância, estendendo sua aplicação além das bibliotecas digitais para documentos disponíveis na web. Dois pontos são fundamentais para aprimorar a precisão da recuperação: (1) esclarecer quais conceitos estão por trás dos termos que ocorrem no texto; (2) determinar novo conceitos semanticamente relacionados aos conceitos encontrados. Mostramos neste artigo quais sequências de palavras formam um termo significativo, ao qual serão aplicados processos de desambiguação baseados no contexto do documento e da vizinhança do termo. Uma vez desambiguado, é determinado um possível contexto espaço-temporal do conceito. Com auxílio de fontes linguísticas da internet, determina-se outros conceitos relacionados semanticamente ao conceito em questão para, em seguida, inserir esta rede de vizinhanças no mapa de tópicos da base de documentos. O tesauro do mapa de tópicos dará suporte à expansão adequada das consultas para determinar, com precisão, os documentos procurados.

Palavras-chave: Recuperação da Informação Semântica. Desambiguação de termos. Indexação. Contexto. Mapa de Tópicos.

Text & Context: towards a more semantic Information Retrieval

ABSTRACT

With the advent of the World Wide Web, Information Retrieval has grown in significance, extending its application beyond digital libraries to documents available on the web. Two points are essential to improve the accuracy of the retrieval: (1) identify the concepts behind the terms that occur in the text; (2) find new concepts semantically related to the concepts found. We show in this article which word sequences form a meaningful term, to which disambiguation processes based on the context of the document and the neighborhood of the term are applied. Once unambiguous, a possible spatio-temporal context of the concept is determined. With the help of linguistic sources on the Internet, other concepts semantically related to the concept in question are determined, and then this neighborhood network is inserted in the topic map of the document base. The topic map thesaurus will support an adequate expansion of a query in order to accurately determine the wanted documents.

***Keywords**: Semantic Information Retrieval. Word sense disambiguation. Document indexing. Context. Topic Map.*

Texto y contexto: hacia una recuperación de información más semántica

RESUMEN

Con el advenimiento de la World Wide Web, la recuperación de información ha ganado importancia, extendiendo su aplicación más allá de las bibliotecas digitales a los documentos disponibles en la web. Dos puntos son esenciales para mejorar la precisión de la recuperación: (1) encontrar qué conceptos están detrás de los términos que aparecen en el texto; (2) determinar nuevos conceptos relacionados semánticamente con los conceptos encontrados. Mostramos en este artículo qué secuencias de palabras forman un término significativo, para aplicar procesos de desambiguación basados en el contexto del documento y la vecindad del término. Una vez desambiguado, se determina un posible contexto espacio-temporal del concepto. Con la ayuda de fuentes lingüísticas en Internet, se determinan otros conceptos relacionados semánticamente con el concepto en cuestión, y luego esta red de vecindad se inserta en el mapa temático de la base documental. El tesauro del mapa de temas apoyará la expansión adecuada de consultas para determinar con precisión los documentos buscados.

Palabras clave*: Recuperación de información semántica. Desambiguación de términos. Indexación. Contexto. Mapa de temas.*

INTRODUÇÃO

Atualmente existem muitas fontes de informação na internet, como dicionários e enciclopédias, que, além de estarem disponíveis universalmente, sofrem contínuas atualizações. Estas fontes também podem ser úteis para programas de indexação de textos que necessitam analisar a correção de escrita de um termo e determinar o significado adequado dele pelo contexto em que ele está situado.

O presente trabalho procura mostrar o que deve ser levado em consideração para identificar todos termos significativos contidos em um texto e encontrar o conceito correto que ele representa.

Para eliminar os 'ruídos' associados aos termos léxicos, terá que ser determinado uma forma canônica para cada termo, eliminando flexões, sinônimos, resolvendo acrônimos e outras irregularidades.

O conceito correto associado a um termo canônico depende do contexto deste termo. Esse contexto é determinado pelo documento em que o termo se encontra e por sua posição nesse documento.

Para cada conceito citado em um texto devemos determinar seu significado preciso (contexto ontológico), um possível tempo específico em que ocorreu (contexto temporal) e um possível lugar (contexto espacial).

Também será dada ênfase na determinação dos conceitos implícitos relacionados com o documento, obtidos pelos relacionamentos semânticos existentes entre conceitos.

Com esta indexação semântica estendida de documentos será possível guiar o pesquisador para determinar com precisão suas necessidades de informação e, com isso, recuperar exatamente os documentos procurados.

ONTOLOGIAS E MAPAS DE TÓPICOS

Enquanto um **dicionário** é um catálogo de termos com a respectiva definição, um **tesauro** é uma rede de termos relacionados semanticamente (MURPHY, 2003). Se um termo possui vários significados, ele terá várias entradas no dicionário, cada uma para um de seus significados. A esta entrada com seu significado chamamos de **conceito**. A fusão de um dicionário de conceitos com as relações semânticas do tesauro correspondente é chamada de **ontologia**.

Enquanto os elementos de uma ontologia são genéricos, i.e., independentes de onde eles ocorrem, os termos de um documento são fortemente ligados ao contexto em que eles aparecem, ou seja, o documento em si e a posição na frase em que ocorrem. Em um processo de indexação de um documento deve-se encontrar os conceitos representados pelos termos contidos nele.

Um **termo** é uma denominação linguística de um ou vários conceitos. Em um documento textual, um termo será formado por uma ou mais palavras. Como um termo pode ser ambíguo, denominando vários conceitos, ele deverá ser desambiguado, associando-se descrições ou conceitos mais genéricos distintos a cada significado dele, obtendo assim, vários termos unívocos. Por exemplo 'jaguar' será {<*jaguar: carro esportivo, modelo*>, <*jaguar: automóvel, fábrica*>, <j*aguar: animal*>}.

Existem dois tipos de ambiguidade de termos. A polissemia e a homonímia. Na polissemia os dois conceitos estão relacionados de alguma forma, podendo um ter sido derivado do outro. É o caso do *jaguar,* em que o modelo de carro é derivado do nome da fábrica de automóveis e estes dois foram derivados do nome do animal. Já na homonímia a ambiguidade é uma pura coincidência da grafia ou pronúncia de dois conceitos. Seria o caso de <*manga: fruta*> e <*manga: camisa*>. Dado um termo *t* utilizamos uma lista de termos ou uma pequena descrição *d* para obter a desambiguação. Chamamos o par *c* = <*t: d*> de **conceito**. Se um termo *t* não é ambíguo denotamos o conceito correspondente por <*t: u*>, em que *u* é uma descrição universal.

Um conceito em um texto pode ter uma validade restrita a certo tempo *tp* e a certa região no espaço *sp*. Assim, a caracterização completa de um conceito será dada por *c* = <*t:(d, tp, sp)*>. Por exemplo em um documento podemos o conceito *c* = <*'Napoleão': (Imperador, 1812, 'Russia')*> se refere a Napoleão quando esteve na Rússia em 1812.

Os conceitos de uma língua podem estar relacionados entre si por relações semânticas, que podem ser hierárquicas, como <*jaguar:animal'* é-um *mamífero*>, <*carro esportivo* é-um *automóvel*> ou <*cauda* parte-de *'jaguar:animal*> ou horizontais, como as relações sinônimo, antônimo, acrônimo e muitas outras. Dado um documento *d*, procura-se indexar este documento encontrando todos os conceitos relacionados a ele.

Com a indexação de um conjunto significativo de documentos será possível, para um conceito *c*, obter todos os documentos que tratam desse assunto e evitar a recuperação de documentos não concernentes. Dado um documento *d* e um conceito *c*, teremos a relação <*c indexa d*>.

Com a ontologia pode-se relacionar não só conceitos contidos em um documento, mas também, conceitos indiretamente associados ao documento. Assim, uma pesquisa por *'carro esportivo'* poderá encontrar documentos que falam de <*jaguar:automóvel*>*"* mesmo que o termo '*carro esportivo'* não ocorra nenhuma vez no documento.

O grafo formado por uma ontologia de um idioma mais um conjunto de documentos com todas as relações *indexa* entre conceitos da ontologia e os documentos é chamado de **Mapa de Tópicos**. A cada relação entre um conceito e um documento poderão estar definidos outros valores, como frequência da ocorrência do conceito e posições onde ele ocorre.

RECUPERAÇÃO DA INFORMAÇÃO

Uma linguagem é formada por um alfabeto, com o qual são construídas palavras, frases, sentenças, parágrafos, documentos e bibliotecas. Um documento descreve algo de interesse para seus leitores. O objetivo central da **Recuperação da Informação (RI)** consiste em, dado um acervo significativo de documentos (uma biblioteca), localizar aqueles que contêm informações de interesse de um pesquisador. Esta necessidade é expressa pela indicação dos assuntos sobre os quais se quer obter os documentos.

O processo da RI pode ser dividido em duas etapas:

a) **Indexação**: Extrair de cada documento novo os assuntos de que ele trata;

b) **Recuperação**: Dar suporte ao usuário para expressar corretamente suas necessidades de informação e apresentar os documentos adequados à consulta.

Neste trabalho são analisados elementos da linguagem natural que podem melhorar a expressividade das informações contidas em um documento. Pretende-se, a partir das palavras e locuções, determinar os conceitos significativos que caracterizam os assuntos contidos no documento. Para os conceitos encontrados são levados em consideração relações semânticas hierárquicas e horizontais que existem entre eles. É construída uma rede semântica de todos os conceitos dos documentos da biblioteca e suas relações semânticas e as associações aos respectivos documentos.

Na recuperação, será essencial guiar o usuário pela rede de conceitos para que ele consiga expressar os conceitos de seu interesse. O objetivo é atender o princípio **TST – Tudo e Somente Tudo**. Ou seja, recuperar todos os documentos de interesse e nenhum que não interesse.

PALAVRAS, TERMOS E CONCEITOS

A indexação de um documento textual parte das palavras contidas nele. A partir das palavras é determinado quais termos significativos elas descrevem e qual conceito cada termo representa.

Veremos nesta seção as questões linguísticas envolvidas nesta determinação dos conceitos representativos de um documento a partir das palavras contidas nele.

PALAVRAS E TERMOS

Um documento em um idioma é uma sequência símbolos válidos naquele idioma como palavras, sinais de pontuação e mais alguns recursos de diagramação.

Para uma palavra ser válida ou ela pertence ao vocabulário definido para este idioma ou ela é uma criação inteligível pela sua posição no texto. Pode ser um nome próprio, uma abreviação ou um termo existente em outro idioma. Cada palavra de um documento tem um significado e irá cumprir um papel específico na sequência de palavras em que está situada.

Várias palavras próximas entre si podem formar um termo significativo para o texto. Assim "*lápis vermelho ou azul*" determina os conceitos de "*lápis vermelho*" e "*lápis azul*". Uma sequência de palavras que expressa um sentido chamamos de **termo**. Um termo formado por uma só palavra é um termo simples, caso contrário ele é um termo composto ou locução. Note-se no exemplo que o term

O uso de termos compostos é essencial para aprimorar a indexação, pois um termo composto pode ter um sentido diferente de seus componentes, possibilitando a remoção de termos desinteressantes para a indexação. Vejamos a palavra "*banco*" que pode ser uma instituição financeira ou um assento. Já o "*banco de dados*" é algo bem diferente. Uma forma bastante útil para identificar um termo composto é utilizar um dicionário. A ocorrência de "*banco de dados*" no dicionário deixa claro que é um termo significativo e, como este conceito não tem nada a ver com os diversos sentidos de "*banco*", o termo "*banco*" não deverá entrar na lista de índices. Se no texto ocorrer "*banco de dados distribuído*", que também é um termo significativo, este deverá ser considerado. Só que, nesse caso, o novo termo é um caso particular de "*banco de dados*", portanto os dois podem entrar na lista.

Além das entradas de um dicionário podem ocorrer diversos termos compostos de interesse para serem escolhidos como índices. Por exemplo, "*banco de dados vazio*" ou "*lápis azul*" do exemplo acima são casos típicos. Estes casos são **substantivos adjetivados** e merecem ser considerados além do substantivo original.

Para determinar os termos significativos baseado nas classes gramaticais dos seus componentes, utiliza-se um analisador morfossintático (POS-Tagger) (BIRD; KLEIN; LOPER, 2009). É um programa que analisa um texto e extrai os termos significativos e determina sua classificação gramatical. Existem muitos termos ou locuções que não interessam para uma indexação. Candidatos a termos chave de um documento são os substantivos e as locuções substantivas (*'banco*', '*casa de saúde*', '*banco de dados*'), eventualmente acompanhados de um adjetivo ('*banco amarelo*'). Outras locuções, como as locuções adjetivas ('*de hoje*'), adverbiais ('*com cuidado*'), conjuntivas ('*desde que*'), prepositivas ('*por meio de*'), pronomiais ('*todo mundo*'), ou verbais ('*querer sair*', '*tem investigado*'), podem ser desconsideradas.

O conjunto de todos os termos significativos extraídos de um documento *d* forma seu **vetor temático** *vt(d)*. A cada termo do vetor ainda pode estar associado a frequência em que ele ocorre no documento. Por exemplo, o texto em inglês:

> *Chapter III: Employing a Grounded Theory Approach for MIS Research*
> *Susan Gasson, Drexel University, USA*
>
> *This chapter provides a brief introduction to the Grounded Theory (GT) approach to research, discussing how it has been used in information systems (IS) research, and how GT studies may be conducted to provide a significant theoretical contribution to the Management Information Systems (MIS) field.*

irá gerar os termos (BISPO, 2012):

> Chapter III(1); Grounded Theory Approach(1); MIS Research(2); Susan Gasson(1); Drexel University(1); USA(1); chapter(5); brief introduction(1); Grounded Theory GT (1); GT(2); approach(1); information systems IS(2); IS(1); GT studies(1); significant theoretical contribution(1); Management Information Systems MIS(1); MIS(2); Field(1)

Como termos muito comuns não são muito expressivos para caracterizar um documento, a importância de um termo para um documento é obtida pelo quociente *tf/idf*, em que *tf* é a frequência do termo no documento e *idf* é a frequência do mesmo termo na coleção completa dos documentos.

Mais detalhes sobre a complexidade de um POS-Tagger podem ser vistos no capítulo 4 da dissertação de Bispo (2012) em que um POS-Tagger foi adaptado para reconhecer adequadamente os termos compostos.

CONCEITOS E CONTEXTOS

Grande parte dos termos (simples ou compostos) contidos em um documento é ambígua, não sendo possível determinar seu significado só a partir da sua ocorrência. Para remover esta ambiguidade deve-se, de alguma forma encontrar o sentido que o termo tem na frase em que ele ocorre.

As coisas significativas que são referenciadas pelos termos de um documento, chamamos de **conceitos**. Podemos distinguir coisas concretas, do mundo físico (*uma pessoa, a chuva de ontem*) e abstratas, do mundo da imaginação humana (*um domínio do conhecimento, um sentimento, uma empresa, um congresso, a noção de chuva, o número 9*) (DORI, 2002).

Existem vários tipos de coisas que merecem figurar como conceitos índice:

Objetos: são coisas individuais (concretas ou abstratas) distintas entre si que têm características específicas e podem ter uma existência delimitada no tempo e no espaço ('*José Bonifácio*', '*meu jaguar*').

Eventos, fenômenos e **processos**: são acontecimentos específicos com uma duração determinada, que delimita um evento (*um congresso, uma aula, uma partida de xadrez*), um fenômeno (*a chuva de ontem*) ou um processo (*a fabricação de um carro*);

Classes ou **tipos** de objetos e eventos: determinam as características comuns a todos objetos ou eventos que se enquadram em seu domínio (*Pessoa, n*úmero inteiro*, Almoço, Chuva, Jaguar: Automóvel*);

Domínios do conhecimento: As diversas ciências e tecnologias, como os itens de um sistema de classificação (*Filosofia. Sociologia, Matemática, Topologia algébrica, Engenharia*, etc.);

Qualidades e **situações**: conceitos que são formados por adjetivos e verbos subjetivados (*roubo, medo, beleza, ideia, progresso, corrupção*).

Frases: trechos mais longos de texto, como poemas, ditados, letras de músicas e outras frases significativas. O sistema deverá manter uma base de frases determinando, para cada frase, detalhes como o autor original, seu uso e outras fontes. Ao varrer o texto a ser indexado, será conferido cada início de uma frase com as frases na base. Havendo uma identificação clara, a frase se tornará um índice.

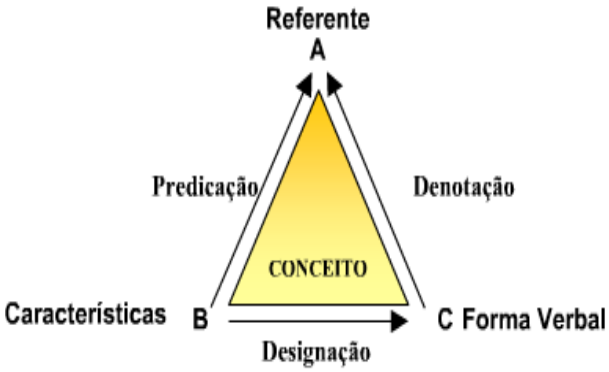
Tanto objetos como eventos são entidades individuais (extensão) ligadas a classes ou tipos (intenção) que determinam ou caracterizam suas instâncias com suas características comuns. Assim '*José Bonifácio*' denomina um objeto que é do tipo *PESSOA* e '*XXX SBBD*' denomina um *SIMPÓSIO*. Denotamos por <'*José Bonifácio*':PESSOA> a pessoa identificada pelo nome José Bonifácio. Como na Wikipedia existem 13 personalidades com esse nome, além do tipo seriam necessários mais elementos para identificar univocamente um objeto. Na ausência de maiores detalhes, supomos que o nome se refere ao *"Patriarca da Independência"* que é o '*José Bonifácio*' mais significativo. Nem sempre é fácil distinguir se um termo em um texto denomina uma instância ou uma classe. Por exemplo em '*um ladrão roubou meu jaguar*' trata-se de um carro específico, enquanto '*o jaguar é um carro esportivo inglês*' refere-se à classe dos carros do modelo Jaguar.

Os domínios do conhecimento, além de também serem candidatos a conceitos da indexação, são especialmente úteis no processo de desambiguação e determinação do contexto de um termo, que será detalhado um pouco mais abaixo.

Para caracterizar claramente um conceito foi sugerida a estrutura denominada de **Triângulo de Dahlberg** (Figura 1), que relaciona uma forma verbal (p.ex. '*José Bonifácio*') que denota um referente (uma pessoa específica com esse nome) com suporte de características/significado ('*o patriarca da independência*').

Segundo a Wikipedia existem mais outros 12 referentes com a mesma forma verbal, mas com outras características. Para nossos propósitos ainda podemos detalhar, nas características, o domínio do referente acrescido de uma descrição. Assim, para a forma verbal `*José Bonifácio*´ temos 13 personalidades referentes distintas e mais 5 localidades referentes.

Figura 1 – Triângulo de Dahlberg



Fonte: Maculan e Lima (2017).

Uma análise detalhada do conceito de "conceito", com todos seus aspectos filosóficos e diversas análises e conceitualizações pode ser vista em Maculan e Lima (2017). Também as variantes e origens do triângulo de Dahlberg (2012) são discutidas neste artigo.

As características de um referente chamamos de seu **contexto**, que deixará claro a que se refere um conceito. Consideramos três contextos que determinam este referente:

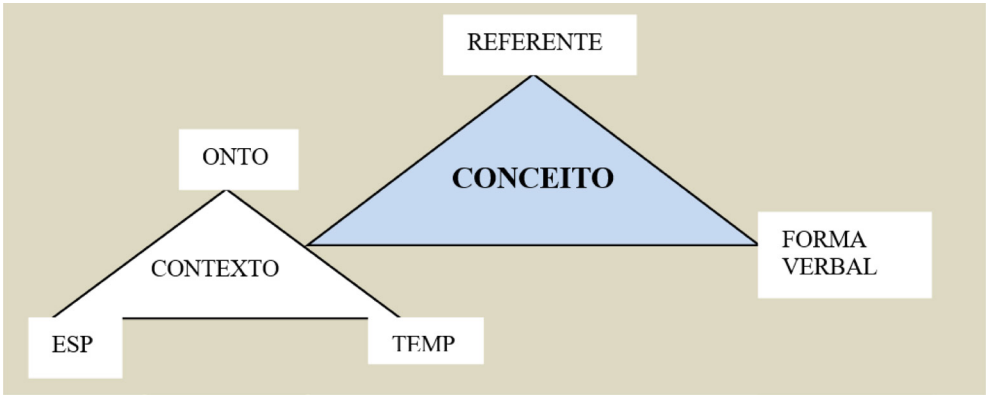
Contexto ontológico: descreve, com precisão, o que o conceito significa, por meio de uma descrição, um domínio ou sinônimos;

Contexto espacial: determina uma possível localização espacial do conceito em questão;

Contexto temporal: determina uma possível ocorrência temporal do conceito em questão.

Para levar em consideração as três dimensões contextuais de um conceito, propomos estender o triângulo de Dahlberg (2012) para forma ilustrada na figura 2.

Figura 2 − Texto e contexto de um conceito



Fonte: Elaborada pelo autor.

O vértice das características (ou contexto) terá três componentes: o ontológico, o espacial e o temporal.

INDEXAÇÃO

A forma mais comum de se referir a conceitos em um documento é por meio de uma linguagem escrita. Esta referência textual pode ser mais ou menos precisa. O objetivo da **indexação** de um documento é encontrar, a partir dos termos contidos nele, os conceitos referenciados pelos termos, também chamados de **assuntos**. A indexação deverá conseguir identificar os assuntos corretos (remover ambiguidades) e identificar formas verbais distintas de um mesmo conceito (sinônimos, anáforas, acrônimos). Este processo cria um vetor temático dos conceitos contidos em um documento. A criação desse vetor passa por 4 etapas:

1) **Localização** e seleção dos termos sintáticos: determinação quais termos são candidatos a assuntos;

2) **Normalização**: encontrar uma forma canônica da escrita dos termos para evitar 'ruídos' como flexões, abreviações, sinônimos e anáforas;

3) **Desambiguação**: Encontrar o significado unívoco de cada termo;

4) **Contextualização espaço-temporal**: determinar um possível contexto espaço-temporal de um conceito;

5) **Criação do vetor temático**: os conceitos identificados são inseridos no vetor. Para cada conceito é contada sua frequência e se tiver um fator tf/idf significativo, é mantido no vetor.

LOCALIZAÇÃO E SELEÇÃO DOS TERMOS

A localização de termos candidatos é obtida por um POS-Tagger[1]. Serão aproveitados os substantivos, locuções substantivas e termos adjetivados. Destes são mantidos os que tiverem um fator tf/idf significativo.

Outra característica fundamental para uma valorização adequada de um termo é sua localização no texto. Se estiver no título geral ou de um capítulo, se estiver no índice remissivo, em uma definição, tudo deve ser levando em consideração. Por exemplo, podemos ter um livro "Álgebra Moderna" em que a palavra 'álgebra' não aparece nenhuma vez no texto, mesmo sendo o termo mais significativo do documento. Ele se destaca por dois motivos: aparece no título e é um termo relacionado diretamente a muitos conceitos importantes contidos no livro, como 'anel', grupo', 'corpo', tudo casos especiais de álgebras.

[1] É um processo da marcação gramatical das palavras simples e compostas de um texto

Para termos compostos nem sempre os componentes também são candidatos. Vejamos a locução "*Banco de Dados Distribuído cheio*". Teríamos os candidatos "*Banco*", "*Dados*", "*Banco de Dados*", "*Banco de Dados Distribuído*" e "*Banco de Dados Distribuído cheio*". Pela análise dos vetores temáticos da definição de cada termo, excluímos os termos "*Banco*" e "*Dados*" por terem uma definição distinta dos termos mais compostos. Já "*Banco* de *Dados*" e "*Banco de Dados Distribuído*" são ambos entradas de um dicionário e o "*Distribuído*" é uma adjetivação de "*Banco* de *Dados*" pois sua definição é derivada dele. Também a locução completa é uma adjetivação de "*Banco* de *Dados Distribuído*", sendo considerado um índice significativo.

NORMALIZAÇÃO

Existem diversas construções linguísticas que podem complicar a determinação adequada dos termos representativos em um texto. Vejamos as principais:

1) **Flexões**: termos no plural devem ser convertidos ao singular. Se o ambiente de indexação exigir a extração não só de termos substantivos, mas também de termos verbais e adjetivos, o número de flexões será bem mais expressivo, incluindo flexões de gênero e conjugações de verbos.

2) **Sinônimos**: escolha do termo mais significativo para representar um conceito. Para termos ambíguos esta escolha terá que ser feita depois da desambiguação, pois um termo terá sinônimos diferentes dependendo do seu significado (*banco ~ assento; banco ~ instituição financeira*).

3) **Figuras de linguagem**: se for detectado o termo da figura. este deve ser eliminado se não for possível aproveitá-lo. Assim na metáfora "*Minha vida era um palco iluminado*" o "*palco iluminado*" só deve ser mantido se identificado como termo metafórico. Outro exemplo seria '*ele é um cachorro*'. A catacrese "*o pé da mesa*" não tem nada a ver com "*pé*". Em metonímias pode-se tentar trocar o termo substituído. Assim "*gosto de ouvir Bach*" leva ao termo "*Música de J.S.Bach*" ou "*comi a caixa de bombons*" leva a "*bombons*" e não a "*caixa de bombons*". Também perífrases e antonomásias poderão ser substituídas (trocar "*país do futebol*" por "*Brasil*"). Na frase "*Alemanha no Campeonato Mundial de Atletismo de 2009*". O termo '*Alemanha*' é uma metonímia de '*Equipe de atletismo da Alemanha*'.

4) **Anáfora**: são termos, como pronomes ou expressões, que se referem a uma entidade citada em um texto, geralmente antes do termo anafórico, que é chamado de antecedente. Existem vários tipos de anáforas: pronomiais (*ele, elas, deles, sua, isto*), verbais ('*esperava ele* ***ganhar*** *no final. Ele* ***conseguiu***') nominais ('***João*** *saiu cedo*. **O *homem*** *não aguentou a conversa*'), adverbiais (('*ele queria me mostrar o* ***jardim***. *Fomos* ***lá***). O antecedente de um termo anafórico pode estar localizado na mesma sentença ou em uma sentença anterior. Na anáfora nominal o termo anafórico nem sempre é um co-referente do antecedente. Pode haver uma relação estrutural (*parte-de, membro-de*) entre os dois. No exemplo do '*João*' acima, temos que '*João*' *membro-de* '*homem*'. O processo de resolução das anáforas consiste basicamente em três etapas: (1) encontrar um termo anafórico, (2) localizar candidatos a antecedentes, (3) escolha do antecedente mais apropriado. Depois disso o termo anafórico deve ser substituído pelo antecedente. Este processo de "resolução de anáforas" *é uma linha de pesquisa bastante ampla, tendo gerado teses*, dissertações e livros (FREITAS, 2005; GODOY, 2010; REINHART, 2016). A tese de Freitas (2005) trata de forma integrada anáforas pronominais e nominais. Para reduzir o número de candidatos a antecedentes, procura primeiro determinar o foco do discurso.

5) **Acrônimos e acrossemia parcial**: Abreviações parciais ou totais de nomes são muito comuns nos mais diversos textos. São siglas de organizações, associações e abreviações de nomes de pessoas. Independente da questão da possível ambiguidade de uma sigla, a ser discutida na próxima subseção, ela deve ser substituída por sua versão por extenso.

Assim, "*John F. Kennedy*", "*J.F. Kennedy*" e "*J.F.K.*" devem todos ser transformados em "*John Fitzgerald Kennedy*". A determinação do significado de um acrônimo muitas vezes está no próprio texto, na primeira citação do acrônimo (p.ex. "*Object Constrait Language (OCL)*") ou pode ser resolvida pelo processo de desambiguação baseada no contexto, citada abaixo.

6) **Palavras substantivadas**: verbos e adjetivos podem ser substantivados. Por exemplo, de "*o carro bonito foi roubado*" podemos extrair os termos "b*eleza*" e "*roubo*".

Muitas das questões de normalização podem ser resolvidas pela consulta a dicionários online como Wordnet[2], Wikipedia[3], Wiktionary[4] ou Infopédia[5].

DESAMBIGUAÇÃO

A desambiguação é uma etapa fundamental para uma indexação adequada de um documento, pois dela depende a determinação adequada do significado do termo, sinônimos, figuras de linguagem e relações semânticas com outros conceitos. Um tratado bastante completo sobre as diversas técnicas sobre esse assunto é o livro de Agirre e Edmonds (2007).

O primeiro passo, para dar suporte à desambiguação dos termos contidos em um documento, é determinar o assunto ou contexto do documento sendo indexado. Para isso, utiliza-se uma base de vetores temáticos dos principais domínios do conhecimento humano. Esta base é criada pelo seguinte algoritmo:

> Para cada domínio do conhecimento como entrada principal da classificação universal CDU[6] (Ciência da Computação, Filosofia, Psicologia, Religião, Ciências Sociais, Matemática, etc.) criar um vetor temático utilizando a descrição desse domínio na Wikipédia. Assim, para cada domínio dom_i obtém-se um vetor $vt(dom_i)$.

Dado um documento *d* é determinado o seu assunto comparando-se, pela função cosseno de Salton e McGill (1983), o vetor temático *vt(d)* com os vetores temáticos dos domínios do conhecimento. Aquele que estiver mais próximo é escolhido como o assunto de *d*. Denotamos esse assunto como *dom(d)*.

Se um termo é composto e não tem registro em um dicionário deve ser reduzido a seu maior componente. Para cada termo *t* ambíguo, há duas possibilidades para desambiguar *t*:

1) Para cada descrição $ds_i(t)$ de um significado de *t*, escolher aquela cujo vetor temático $vt(ds_i(t))$ esteja mais próximo de *vt(dom(d))*.

2) Escolha a 'vizinhança' de *t* dada pela frase ou o parágrafo em que *t* está contido, denotada por *viz(t)*, e escolha o $vt(ds_i(t))$ mais próximo de *vt(viz(t))*.

O resultado desta escolha é denotada por *dom(t)*. Caso haja divergência entre os dois resultados deve ser dada preferência ao critério da vizinhança, que dá mais ênfase ao significado local do termo. A primeira opção deve ser escolhida nos casos em que a vizinhança não for suficientemente expressiva. Como resultado desse processo a cada termo ambíguo *t* é associado seu significado *dom(t)* obtendo-se *t:dom(t)*.

O uso dos 10 domínios principais de uma classificação universal como a CDU como critérios de desambiguação pode ser insuficiente para remover todas as ambiguidades. Como no caso do "*Jaguar:Automóvel*" e "*Jaguar:Indústria*" seriam classificados como a classe CDU *6-Tecnologia*. Os domínios deverão ser detalhados mais até que se consiga a desambiguação total. No exemplo, usaríamos *656-Transporte* e *67-Indústria*, sabendo-se que *Automóvel* é uma especialização de *Transporte*.

Um caso particular de ambiguidade pode ocorre com os acrônimos. Por exemplo, o acrônimo *OCL* contempla, na Wikipedia em francês, três significados: *Object Constraint Language*, *Organisation Communiste Libertáire* e *Orchestre de Chambre de Lausanne*.

[2] http://wordnetweb.princeton.edu/perl/webwn

[3] https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal

[4] https://pt.wiktionary.org/wiki/Wikcion%C3%A1rio:P%C3%A1gina_principal

[5] https://www.infopedia.pt/dicionarios/lingua-portuguesa

[6] Usamos aqui a Classificação Decimal Universal (CDU), mas também há outras classificações como, por exemplo, a IDC – Information Coding Classification (DAHLBERG 2012)

Cada significado da sigla é usado como descrição e é testado pelo seu contexto, sendo selecionado o mais adequado. Como termo resultante usa-se a versão por extenso combinada com a sigla, no exemplo poderia ser *Object Constraint Language (OCL*).

Um caso mais complicado é a ambiguidade de uma frase. Por exemplo, em '*o homem viu a moça com o telescópio*', não está claro quem está com o telescópio.

A determinação do significado preciso de um termo, denominamos sua **contextualização odontológica**.

CONTEXTUALIZAÇÃO ESPAÇO-TEMPORAL

Muitos objetos ou eventos têm contextos espaciais onde e quando existiram ou aconteceram. Por exemplo, podemos nos interessar em documentos que relatam algo sobre Napoleão na Rússia ou sobre o Governo dos 100 dias de Napoleão após seu retorno do exílio em Elba. Ou seja, para documentos que tratam de um objeto como o imperador Napoleão, pode ser interessante determinar as coordenadas espaciais e temporais da citação deste objeto no texto garantindo um contexto mais aprimorado desta ocorrência.

Os dois tipos de coordenadas podem ser obtidos em dois passos incorporados à indexação de um documento:

1) Detecção dos objetos espaciais e temporais contidos no documento. Nos exemplos, seria encontrado '*Rússia*' como um objeto espacial e '*Governo dos 100 dias*' como um objeto temporal, determinando-se valores para esp(*Rússia*)=< 60 00 N, 100 00 E> e temp(*Governo dos 100 dias*)=<26/02/1815, 15/10/1815>. Entre os objetos temporais distingue-se tempos explícitos '*1812*', implícitos '*Governo dos 100 dias*' e relativos '*5 dias depois*' (SCHILDER, F.; HABEL, 2005). Se *t* é um tempo implícito ou relativo, denotamos *temp(t)* seu respectivo tempo explícito. Se o tempo relativo não é preciso (p.ex. '*depois da batalha*') é denotado por *dep(t)* para 'depois' e *ant(t)* para 'antes'. Analogamente consideramos *esp(o)* o espaço de um objeto *o*. Se *o* é um objeto móvel, consideramos esp(*o,t*) o espaço em que *o* esteve durante o tempo *t*. Por exemplo, podemos ter esp(*'Napoleão', 1812*). Inversamente, podemos ter temp('*Napoleão', esp(Russia)*).
2) Com todos os objetos espaciais e temporais identificados, os valores desses objetos podem ser associados aos objetos existentes nas suas vizinhanças. Se possível, o documento como um todo também pode receber estas coordenadas. É claro que a associação só faz sentido para objetos adequados, que têm uma mobilidade no espaço e no tempo.

A definição destas coordenadas para um conceito chamamos de **contextualização espaço-temporal**.

Um termo com suas contextualizações completas terá a forma:

Termo: (*domínio, espaço, tempo)*

Os dois exemplos citados ficariam

'Napoleão':(imperador, < 60 00 N, 100 00 E>*,{})*

'Napoleão':(imperador, {}, <26/02/1815, 15/10/1815>*)*

Se o contexto espacial ou temporal não estiver explícito, como nos dois exemplos citados, ele poderá ser obtido de forma implícita. Por exemplo na frase "*Alemanha no Campeonato Mundial de Atletismo de 2009*". Após a metonímia '*Alemanha*' ter sido substituída por '*Equipe de atletismo da Alemanha*' a fonte BabelNet fornecerá os contextos espaciais e temporais, determinando o conceito final:

'*Equipe de atletismo da Alemanha':({}, esp(Berlim), <15/08/2009, 25/08/2009>)*

Vale destacar que os contextos espaciais e temporais devem ser aproveitados adequadamente. Por exemplo, tanto consultas por documentos sobre Napoleão em Moscou como sobre Napoleão na Ásia devem retornar o documento do primeiro exemplo.

Consultas a documentos sobre Napoleão durante a batalha de Waterloo devem retornar o segundo documento, já que *temp(Batalha de Waterloo)* = *<18/16/1815>* ⊂ *<26/02/1815, 15/10/1815>*.

Além de considerar os contextos espaciais e/ou temporais de um conceito, estes contextos por si só também podem ser objeto de uma consulta. Por exemplo, pode ser feita uma consulta a documentos que discorrem sobre coisas ou ocorrências no espaço da Rússia no ano de 1812.

RELAÇÕES SEMÂNTICAS, TESAUROS E MAPAS DE TÓPICOS

Em um ambiente de recuperação da informação, nem sempre os conceitos explícitos em um documento são suficientes para caracterizá-lo adequadamente. Pode-se fazer uma pesquisa por '*animal doméstico*' e espero receber documentos que falam de '*cachorro*' e/ou de '*gato*', sem necessariamente conterem a palavra '*animal doméstico*'. Pode-se pesquisar por '*peixe*' e um documento trata de '*cardume*' ou pesquisar por '*mão*' e o documento tratar de '*dedo*'. Também posso querer pesquisar por documentos que tratam dos *inimigos* de Napoleão.

Para considerar situações semelhantes às exemplificadas, devemos levar em consideração possíveis relações semânticas existentes entre os conceitos de um dicionário.

Um sistema de organização do conhecimento contempla três concepções: conceito, termo e relacionamento. Melo e Bräscher (2014) analisam esta forma de organizar o conhecimento segundo dois pontos de vista filosóficos distintos: O positivismo e o pragmatismo. Segundo o positivismo, estudado por Dahlberg (2012), o conhecimento é algo mais universal e estático, enquanto no pragmatismo, defendido por Hjorland (2009), ele é mais dinâmico e contextual. Com a contextualização ontológica, espacial e temporal proposta aqui, acredita-se poder atender aos requisitos pragmáticos dos conceitos que, com isso, têm sua validade limitada a restrições espaciais e/ou temporais.

Relações semânticas também podem ter um contexto espaço-temporal para determinar sua validade. Por exemplo, duas pessoas A e B podem ser *parceiros* em um time de futebol e *adversários* em uma partida de tênis. Ainda mais, no emprego B pode ser *empregado-de* A. Uma relação *casado-com* terá um contexto temporal no qual ela é ou foi válida.

Também há relações que podemos chamar de **atributivas**, pois conferem qualidades a um objeto. Assim a idade, o endereço e outros dados de uma pessoa, seriam relações deste tipo.

As relações semânticas a serem detalhadas são relações estruturais e universais que independem do contexto e são encontradas em dicionários/tesauros como a WordNet.

Podemos classificar as relações em dois tipos: Verticais ou hierárquicas e horizontais.

RELAÇÕES VERTICAIS OU HIERÁRQUICAS

São relações entre dois conceitos em que um é mais abrangente ou genérico e o outro mais específico ou detalhado. As principais são:

Hiperônimo/Hipônimo: Esta relação também é conhecida como generalização/especialização ou relação é-um. Um objeto ou conceito mais genérico é reconsiderado a um nível mais específico, por exemplo *rosa* < *flor*. Usamos a notação '<' para descrever esta relação. Se a relação se dá entre classes de objetos, falamos em subclasse e classe, como *automóve*l < *ve*ículo-*de-transporte* ou *imperador* < *pessoa*. Um conceito também pode ter vários hiperônimos, como *gato* < *animal-caseiro* e *gato* < *felino*. Esta relação é transitiva pois de *jaguar* < *automóvel* também vale *jaguar* < *veículo de transporte*. Também entram nessa categoria objetos com restrições espaço-temporais. Por exemplo, vale *'Napoleão':(imperador, esp(Russia),{})* < *'Napoleão':(imperador, {},{})*, ou seja o Napoleão na Rússia é um hipônimo de Napoleão como um todo. Um objeto com um contexto espacial e/ou temporal pode ser considerado um hipônimo dele se este contexto.

Classificação/Instanciação: Neste caso temos a relação conhecida como *instância-de*. Ela relaciona um objeto ou evento com sua classe ou tipo. Denotamos esta relação com '<<'. Temos, por exemplo, '*Napoleão*':*imperador* << *pessoa* ou *meu-carro* << *jaguar*. Em certos contextos uma classe pode ser instância de uma metaclasse. Assim, se *pr* é um programa em Java, temos *pr*<<*Programa-Java*<<*Linguagem-de-programação*.

Holonímia/Meronímia: É a relação entre um objeto composto e suas partes, conhecida como agregação ou relação *parte-de*. Com a notação '∠' teríamos *cabeça*∠*corpo* ou *morfema*∠*palavra*. Aqui também vale a transitividade, pois de *olho*∠*cabeça* e *cabeça*∠*corpo* obtemos *olho*∠*corpo*. Um caso especial de meronímia é a composição homeômera ou relação *membro-de*. É quando um objeto é um grupo de elementos do mesmo tipo. Sería o caso de *estudante*∠*classe*. Relações atributivas também podem ser consideradas meronímias. Assim, se p1 é uma pessoa de nome Nicolas, podemos considerar 'Nicolas'∠p1.

Para estas relações existem propriedades que também podem ser úteis. Dados objetos x, y e z valem as regas:

$x < y$ e $y < z \Rightarrow x < z$ (transitividade da generalização)

$x \angle y$ e $y \angle z \Rightarrow x \angle z$ (transitividade da agregação)

Se $x \angle z$ e $y < z \Rightarrow x \angle y$ (a estrutura de uma classe é herdada para suas subclasses);

$x << y$ e $y < z \Rightarrow x << z$ (instâncias de uma subclasse são instâncias da superclasse)

RELAÇÕES HORIZONTAIS OU ASSOCIAÇÕES

Objetos de um mesmo nível (de extensão ou de intenção) podem ter uma relação especial entre si que merece ser considerada. Podemos distinguir relações linguísticas e conceituais.

Não incluímos nas relações horizontais relações linguísticas gramaticais, como flexões de uma palavra, ou relações fonéticas, como homofonia ou paronímia. Estas ou não tem importância ou foram resolvidas no processo de normalização dos termos.

Relações linguísticas:

Sinonímia: quando dois termos têm um mesmo significado. Esta qualidade é muito sensível ao contexto. Por exemplo, a palavra *alto* pode ser sinônimo de *agudo* em um contexto e de *grande* em outro. Denotamos a sinonímia entre x e y como *x~y*. Vemos que vale se *x~y* então *y~x*. Mas, como a sinonímia nem sempre é absoluta, a transitividade pode não valer e *x~y* e *x~z* pode não implicar em *y~z*.

Antonímia: um termo 'x' é um antônimos de 'y' quando uma ocorrência de 'x' em uma frase pode ser substituída por 'não y'. Denotamos antonímia por *x*||*y*. Há antônimos contraditórios (*vivo*||*morto*), contrários parciais (*quente*||*morno*||*frio*) ou direcionais (*mãe-de*||*filho-de* ou *entrada*||*saída*). Esta relação é reflexiva, pois também valerá 'y'||'x';

Co-hipônimos: duas subclasses de uma mesma superclasse são classes irmãs. Por exemplo, *gato*<*co-hip*>*cachorro*, pois *gato* < *animal-doméstico* e *cachorro* < *animal-doméstico*.

Autohipônimo: uso de um hiperônimo como hipônimo. Uso de *cachorro* no lugar de *cachorro macho*.

Relações conceituais:

Instancia: ao contrário da relação '*instância-de*' que relaciona um objeto a sua classe, esta relação associa um objeto concreto a sua versão abstrata. Por exemplo, temos *exemplar-de-livro*<*instancia*>*livro*. Ele terá todos os atributos do objeto abstrato mais detalhes sobre sua existência física, como localização e estado;

Refina: este relacionamento acrescenta mais

detalhes a um objeto mais genérico. Por exemplo *conteúdo<refina>ficha-de-livro*.

Indexa e deriva: Do conteúdo de um livro derivamos as palavras-chave que o indexam. Teríamos, então, as relações *conteúdo<refina>livro; conteúdo<deriva>palavra-chave* e *palavra-chave<indexa>livro*.

Outras: além das relações citadas, outras poderão ser definidas como *ama, amigo-de, chefia, toca*, etc.

Vale ressaltar que as relações horizontais também podem ser divididas em homeômeras (do mesmo tipo) e heterômeras (tipos diferentes). Sinônimos, antônimos e co-hipônimos são homeômeras enquanto as relações conceituais (instancia, refina, indexa) são heterômeras. As relações homeômeras costumam ser simétricas e podem ser reflexivas e transitivas. Já uma relação heterômera sempre será assimétrica.

Para todas as relações semânticas, pode ser interessante determinar quais regras se aplicam a cada uma delas, como reflexividade, transitividade, simetria e outras, além de possíveis regras inter-relacionamentos. Uma análise formal das relações de hiperônimo e merônimo foi feita por Schiel (1989).

Mais detalhes sobre as principais relações semânticas aqui discutidas podem vistos no livro de Murphy (2003). Uma taxonomia destas relações é dada por Maia, Lima e Maculan (2017).

INTEGRAÇÃO AO TESAURO

Na seção anterior foram descritas as ações necessárias para converter um termo encontrado em um documento no conceito que ele representa. Considerando o mapa de tópicos já existente, construído pelo processamento de documentos anteriores, cada novo conceito deve ser integrado a ele. Se ele a existir no tesauro só será acrescida a relação deste termo com o documento sendo indexado. A esta relação pode estar associado um fator de relevância da ocorrência do conceito no documento assim como os locais onde o respectivo termo aparece.

Caso o conceito seja novo, serão pesquisadas as possíveis relações semânticas com os conceitos existentes. Em caso afirmativo, o conceito principal será conectado pela respectiva relação semântica ao conceito existente. Conceitos novos formarão novos nós no grafo do tesauro, conectados aos nós existentes pelas relações semânticas encontradas.

Com todos os conceitos encontrados integrados ao tesauro existente e relacionados ao documento correspondente, temos uma versão atualizada do Mapa de Tópicos.

CONSULTAS

Todo o trabalho de indexação semântica descrito nas seções anteriores tem por objetivo possibilitar a formulação de consultas precisas sobre a necessidade de obter fontes adequadas sobre um a assunto bem especificado.

Tem-se a disposição uma rede semântica de conceitos bem contextualizados nas dimensões ontológica, espacial e temporal, interligados entre si pelas mais diversas relações semânticas, gramaticais, estruturais. Cada conceito foi obtido ou diretamente de um documento durante sua indexação ou foi acrescentado pelas consultas a termos semanticamente relacionados descritos nos dicionários e enciclopédias acessados.

Acoplado a essa rede está a base de documentos, cada um associado a todos os conceitos básicos que são tratados nele.

O processo de criação da consulta é composto por duas etapas:

1) Determinação dos conceitos: Para montar sua consulta o usuário especifica, como em um sistema convencional de RI, os termos significativos. Para cada termo é verificada sua ou suas ocorrências como denominação de conceitos no tesauro. De cada conceito é mostrado ao usuário seu contexto ontológico (o significado). O usuário escolhe o contexto

ontológico desejado, e pode definir possíveis coordenadas espaço-temporais. Do conceito ontológico '*Napoleão*';(imperador) ele pode designar o espaço, p.ex. *esp(França)* e o tempo, p.ex. '<1820, 1821>'. Todos os conceitos que interceptam estes parâmetros deverão ser considerados. Também poderá ser definido um limite de frequência relativa do termo ou detalhes sobre sua localização no documento (p.ex. no título).

2) Expansão da consulta: pela posição do conceito no tesauro, deverão ser oferecidas possíveis expansões da consulta pelas relações semânticas. No caso no '*Napoleão*';(imperador) poderíamos ter as opções: *Genérico:* □ *'Imperador',* □ *"general francês'; Co-hipônimo:* □ *'Alexandre I':(czar da Rússia),* □ *'Louis Bonaparte':(Rei da Holanda),* □ *'Francisco II':(Imperador Romano-Germânico);* A consulta também pode ser expandida para outros significados de '*Napoleão*'. Assim poderá ser acrescido o conceito □ *'Napoléon':(Conhaque*). Também pode-se escolher, propositalmente, uma metáfora. Assim, p. ex., queremos documentos em que '*matar*' é usado em sentido figurado, como '*matar a fome*' ou '*matar o tempo*'. Todos conceitos de uma pesquisa poderão ser combinados por conexões $\vee$ (ou) e $\wedge$ (e).

É claro que, após a execução de uma consulta, ela pode ser refinada, alterando os parâmetros dos termos.

TRABALHOS RELACIONADOS

Como o ambiente aqui proposto envolve diversas etapas, como desambiguação de termos, vários tipos de contextualização e expansão por relações semânticas, há trabalhos relacionados específicos para estas etapas.

Loh, Wives e Oliveira (2000) desenvolveram um trabalho de descoberta de conhecimento em textos (KDT – Knowledge Discovery in Texts) que pretende descobrir os conceitos contidos em textos. Só que eles não usam dicionários públicos mas, antes da pesquisa nos textos, é determinado o domínio dos documentos e potenciais conceitos. Para cada termo extraído é analisado se há conceitos candidatos.

Mihalcea (2007) mostra a utilidade do uso de uma combinação das desambiguações da Wikipédia, utilizando explicações mais precisas dos conceitos no WordNet, para obter uma boa desambiguação de termos.

Um sistema bastante maduro de uma ontologia de termos é o WordNet. Para cada termo léxico mostra seus diversos significados com os respectivos sinônimos, hipônimos, hiperônimos e vários outros tipos de termos relacionados. Cada caso é acompanhado de uma explicação e exemplos. Infelizmente, com este detalhamento só existe a versão em inglês. Foi projetada uma versão multilíngue, o EuroWordNet[7], mas o projeto foi abandonado em 1999 e também não teve uma versão em português. Para o português um projeto semelhante é o Onto.PT[8] que foi construído a partir de vários Thesaurus, Dicionários, Enciclopédias e Corpora em português (OLIVEIRA; GOMES 2014).

Um sistema de uma rede semântica multilíngue com um grande número de relações semânticas é a BabelNet[9]. Entre as coisas representadas por termos é feito distinção entre diversas categorias como conceitos, nomes, esportes, literatura, etc.

A BabelNet oferece uma opção de visualização da rede vizinha a um conceito e exposição, a partir do conceito original, de novaos nós iterativamente. Apesar de possuir um número muito expressivo de relações semânticas quase todas elas são identificadas só pelo nome '*relacionado semanticamente*', sem maiores especificações.

A rede que serve de suporte às diversas redes da família Wikimedia é a Wikidata[10]. Ela é bastante completa, cada entrada possui um identificador único, uma descrição, relações padrão como 'instância de', 'subclasse de' e muitas relações

[7] http://projects.illc.uva.nl/EuroWordNet/
[8] http://ontopt.dei.uc.pt/
[9] https://babelnet.org/
[10] https://www.wikidata.org/wiki/Wikidata:Main_Page

específicas do tipo da entidade. Por exemplo, para '*Jaguar*' temos as entradas:

Onça-pintada (Q35694): felino americano

instância de (*Taxon*); subclasse de (*mamíferos*; nome comum (*Jaguar*); idade máxima observada (*28 anos*); distribuição geográfica (*América central e América do Sul*);

Jaguar (Q26742231): marca de carros inglesa da Jaguar Land Rover

Instância de (*marca/fabricante de carro*)

Jaguar Land Rover (Q6122893) Fabricante britânica de automóveis

Instância de (*montadora de automóveis*); data da fundação (*1 janeiro 2013*); proprietário (*Tata Motors*); proprietário de (*Jaguar* (Q26742231), *Land Rover*); Identificadores (...)

SS Jaguar (Q7393881): Página de desambiguação da Wikimedia

Jaguar (cartunista) (Q6122865): cartunista brasileiro

Instância de (ser humano); sexo ou gênero (masculino);

Note-se que uma simples palavra pode designar um tipo de animal, uma pessoa concreta ou uma página na internet.

A dissertação de Hernani Costa (2010) propõe um sistema bastante amplo para a extração de ontologias de textos em português. Ele analisa a semântica das frases e determina várias relações semânticas.

Para a caracterização temporal de textos existe uma atividade extensa nos últimos 10 anos, quando foi elaborada uma proposta de avaliação de sistemas de indexação temporal denominada TempEval durante o Workshop SemEval 2007 (VERHAGEN *et al.,* 2007). Esta proposta de critérios evoluiu com o tempo para a versão TempEval-3 em 2014 (UZZAMAN *et al.*, 2014). O objetivo de uma anotação temporal consiste em associar eventos tanto a expressões temporais no texto como ao tempo de criação do documento e, além disso, ordenar os eventos citados no documento.

Apesar de existir um interesse grande na consideração de aspectos temporais associados à Recuperação da Informação[11], não encontramos nenhum trabalho que foca na determinação do contexto temporal de um conceito segundo os termos temporais na sua vizinhança.

CONCLUSÃO

O presente trabalho apresenta uma gama de ideias com o intuito de servir para melhorar a qualidade das informações recuperadas em um processo de RI. É claro que não é fácil ou até impossível automatizar todas as sugestões apresentadas. Sabemos como é subjetivo o tratamento e a interpretação adequada da linguagem natural. A prova disso são os tradutores automáticos que, até hoje, ainda apresentam sérias limitações.

Um dos problemas é a identificação adequada dos conceitos contidos em um documento. Mas, acreditamos que esta dificuldade poderá ser reduzida com o contínuo acréscimo de novos documentos. Ou seja, o esforço de intervenção manual na classificação dos termos, irá se reduzir gradativamente com o crescimento do Mapa de Tópicos, em constante evolução.

Uma parte considerável das tarefas aqui propostas, já foi desenvolvida no âmbito de um projeto academico de pesquisa, denominado **RISO – Recuperação da Informação Semântica de Objetos Textuais**[12].

A primeira etapa de criação de vetores temáticos dos domínios do conhecimento e extração dos termos de um documento foi objeto da dissertação de mestrado de Bispo (2012). A desambiguação de termos e a determinação dos conceitos designados foi trabalhada por ARAÚJO JÚNIOR (2013) e Araújo

[11] https://en.wikipedia.org/wiki/Temporal_information_retrieval#References

[12] https://sites.google.com/a/copin.ufcg.edu.br/riso-t/home

Júnior; Schiel; Marinho (2015). Para determinar o contexto temporal dos conceitos, primeiro tiveram que ser extraídos os termos temporais como datas, anos e outras formas (SANTOS, 2013; SANTOS; SCHIEL, 2013).

Com as expressões temporais identificadas pode-se definir o contexto temporal dos conceitos próximos a eles (ALVES, 2016). O contexto espacial ainda não foi integrado ao projeto RISO, mas esta integração poderia ser realizada utilizando-se o motor de busca geográfica GeoSEn, objeto da dissertação de mestrado de Cláudio Campelo (CAMPELO; BAPTISTA, 2009).

O processamento de consultas utilizando a estrutura de Mapa de Tópicos criado na indexação é descrita em Schiel *et.al.* (2014). A consulta passa pelas etapas de desambiguação, expansão e execução.

REFERÊNCIAS

AGIRRE, E.; EDMONDS, P. (eds.). *Word Sense Disambiguation:* algorithms and applications. Netherlands: Springer, 2007. DOI 10.1007/978-1-4020-4809-8.

ALVES, G. M. R. *RISO-GCT:* Determinação do Contexto Temporal de Conceitos em Textos. 2016. 95 f. Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Paraíba, Brasil, 2016. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/469. Acesso em: jan. 2021.

ARAÚJO JÚNIOR, J. G. *Uma abordagem para a Indexação Semântica de Documentos Textuais baseada em Fontes Heterogêneas de Informação*. 2013. Dissertação (Mestrado em Ciência da Computação) – Pós-Graduação em Ciência da Computação, Universidade Federal de Campina Grande, Campina Grande, Paraíba, 2013. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/4878. Acesso em: jan. 2021.

ARAÚJO JÚNIOR, J. G.; SCHIEL, U.; MARINHO, L. B. An approach for building lexical-semantic resources based on heterogeneous information sources. *In:* ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 30., 2015, Salamanca. *Proceedings* [...]. Spain: SAC'15, Apr. 2015. Disponível em: https://doi.org/10.1145/2695664.2695896. Acesso em: jan. 2021.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python:* analyzing text with the Natural Language Toolkit. 1st ed. [*S.l*]: O'Reilly, 2009. Disponível em: https://www.nltk.org/book/. Acesso em: 1 jan. 2021

BISPO, M. C. T. *Criação de vetores temáticos de domínios para a desambiguação polissêmica de termos.* 2012. 100f. Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-graduação em Ciência da Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Paraíba, 2012. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/1314. Acesso em: fev. 2021.

CAMPELO, C.; BAPTISTA, C. *A Model for Geographic Knowledge Extraction on Web Documents*. *In:* INTERNATIONAL WORKSHOP ON SEMANTIC AND CONCEPTUAL ISSUES IN GIS, 3., 2009. *Proceedings* […]. Gramado: SeCoGIS, 2009. p. 317-326.

COSTA, H. Automatic Extraction and Validation of Lexical Ontologies from text, Diss. Mestrado, Universidade de Coimbra, Portugal - September, 2010, 124 páginas

DAHLBERG, I. A Systematic New Lexicon of All Knowledge Fields based on the Information Coding Classification. *Knowledge Organization,* [*s.l.*], v. 39, n. 2, p. 142-150, 2012. Disponível em: https://www.ergon-verlag.de/isko_ko/downloads/ko_39_2012_2_j.pdf. Acesso em: fev. 2021.

DORI, D. *Object-Process Methodology*. Berlin: Springer Verlag, 2002.

FREITAS, S. A. A. *Interpretação Automatizada de Textos:* processamento de Anáforas. 2005. 184 f. Tese (Doutorado em Engenharia Elétrica) - Programa de Pós-Graduação em Engenharia Elétrica, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2005. Disponível em: http://repositorio.ufes.br/handle/10/4114. Acesso em: fev. 2021.

GODOY, M. C. *Resolvendo a Anáfora Conceitual:* um Olhar para além da Relação Antecedente/ Anafórico. 2010. 78 f. Dissertação (Mestrado) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, SP, 2010. Disponível em: http://repositorio.unicamp.br/jspui/handle/REPOSIP/269041. Acesso em: ago. 2021.

HJORLAND, B. Concept theory. *Journal of the American Society for Information Science and Technology*, v. 60, n. 8, p. 1519-1536, 2009

LOH, S.; WIVES, L.K.; OLIVEIRA, J.P. de *Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva*, INTERNATIONAL SYMPOSIUM ON KNOWLEDGE MANAGEMENT/DOCUMENT MANAGEMENT (ISKM/DM, Curitiba, p. 125-147, 2000

MACULAN, B. C. M. S.; LIMA, G. A. B. O. Buscando uma definição para o conceito de 'conceito'. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 22, n. 2, p. 54-87, jan./abr. 2017. DOI https://doi.org/10.1590/1981-5344/2963. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2963. Acesso em: jan. 2021

MAIA, L. S.; LIMA, G. A. B. O.; MACULAN, B. C. M. S. Taxonomia dos tipos de relações semânticas para a organização e a representação do conhecimento: uma proposta a partir da literatura. *Tendências da Pesquisa brasileira em Ciência da Informação*, [*s.l.*], v. 10, n. 2, ago./dez. 2017. Disponível em: https://revistas.ancib.org/index.php/tpbci/article/view/419. Acesso em: fev. 2021.

MELO, M.A.F.; BRÄSCHER, M. Termo, conceito e relações conceituais: um estudo das propostas de Dahlberg e Hjorland. *Ciência da Informação,* v. 43 n. 1 p. 67-80 jan./abr., 2014

MIHALCEA, R. Using Wikipedia for Automatic Word Sense Disambiguation. *In:* NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007, New York. *Proceedings* […]. New York: Association for Computational Linguistics, Apr. 2007. Disponível em: https://aclanthology.org/N07-1025/. Acesso em: fev. 2021.

MURPHY, L M. *Semantic Relations and the Lexicon*. USA: Cambridge University Press, 2003. Disponível em: https://doi.org/10.1017/CBO9780511486494. Acesso em: ago. 2021

OLIVEIRA, H. G.; GOMES. P. Onto.PT: recent developments of a large public domain portuguese wordnet. *In:* GLOBAL WORDNET CONFERENCE, 7., 2014, Estonia. *Proceedings* […]. Estonia: University of Tartu Press, Jan. 2014. p. 16-22. Disponível em: https://aclanthology.org/W14-0103/. Acesso em: fev. 2021.

REINHART, T. *Anaphora and semantic interpretation*. London: Routledge Library Editions, 2016. Disponível em: https://doi.org/10.4324/9781315536965. Acesso em: jan. 2021.

SALTON, G.; MCGILL, M. J. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill, 1983.

SANTOS A. A. *RISO-TT:* Extração de Expressões Temporais em Textos. 2013. 84 f. Dissertação (Mestrado em Ciência da Computação) - Pós-graduação em Ciência da Computação, da Universidade Federal de Campina Grande, Paraíba, 2013. Disponível em: http://dspace.sti.ufcg.edu.br:8080/xmlui/handle/riufcg/1263. Acesso em: mar. 2021.

SANTOS, A. A.; SCHIEL, U. *Recovery of Temporal Expressions from Text:* The RISO-TT Approach. *In:* INTERNATIONAL CONFERENCE OF SEMANTIC PROCESSING, 2013, Lisbon. *Proceedings* […]. Portugal: SEMAPRO, 2013.

SCHIEL U. Abstractions in Semantic Networks: Axiom Schemata for Generalization, Aggregation and Grouping, *ACM-SIGART Bulletin* No. 107, p. 25-26, 1989

SCHIEL, U. *et al. Semantic Information Retrieval considering Term Disambiguation and Linguistic Enrichment*. Departamento de Sistemas e Computação, Universidade Federal de Campina Grande, Paraíba, 14 jan. 2014. Relatório técnico.

SCHILDER, F.; HABEL. C. From temporal expressions to temporal information: semantic tagging of news messages. *In:* MANI, I.; PUSTEJOVSKY, J.; GAIZAUSKAS, R. (ed.). *The Language of Time:* a Reader. Oxford, UK: Oxford University Press, 2005. p. 533–544.

UZZAMAN, *et al.* SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events and Temporal Relations. *In:* INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATIONS, 7., 2013, Atlanta, Georgia. *Proceedings* [...]. Atlanta, Georgia: Association for Computational Linguistics, 2013. p. 1–9. Disponível em: https://aclanthology.org/S13-2001/. Acesso em: fev. 2021.

VERNHAGEN, M. *et al.* SemEval-2007 Task 15: TempEval Temporal Relation Identification. *In:* INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATIONS, 4., 2007, Prague, Czech Republic. *Proceedings [...].* Prague: Association for Computational Linguistics, 2007. Disponível em: https://aclanthology.org/S07-1014. Acesso em: jan. 2021.