# Identificação da Cobertura Espacial de Documentos usando Mineração de Textos

**Rosa Nathalie Portugal Vargas**[1], **Maria Fernanda Moura**[2], **Eduardo Antonio Speranza** [2], **Solange Oliveira Rezende** [1]

[1]Instituto de Ciências Matemáticas e Computação – Universidade de São Paulo (ICMC-USP) São Carlos – SP – Brasil

[2]Embrapa Informática Agropecuária Campinas, SP – Brasil

{nathalie,solange}@icmc.usp.br

{fernanda,speranza}@cnptia.embrapa.br

***Abstract.*** *Nowadays, it is usual that users take into account the geographical context of the documents in the Retrieval Information process. However, the conventional retrieval information systems based on key-word matching do not consider which words can represent geographical entities that are spatially related to other entities in the documents. To solve this problem, it is necessary to enable the geo-referencing of texts by identifying the geographical entities present in the text and associate them with their correct spatial location. Thus, the main strategy to overcome these problems include the identification of evidence to assist in the identification and disambiguation of places in the texts. This work proposes a methodology that allows the identification and spatial localization of the documents, denominated SpatialCIM. The SpatialCIM methodology has the objective to organize the Topônym Resolution process.We proposed and developed the approaches of (1) Disambiguation for Points and (2) Textual and Structural Disambiguation. These approaches exploit two different techniques of toponym disambiguation, which generate and desambiguate the associated paths with the recognized geographical toponym for each document. From the results it was possible to demonstrate that the disambiguation techniques improve the precision and recall for the spatial classification of documents. The positive effect of using a linguistic tool for the process of geographical entities recognition was also demonstrated. Thus, the usefulness of the disambiguation process for obtaining a spatial coverage of the documents was proved.*

***Resumo.*** *Atualmente, é comum que usuários levem em consideração o contexto geográfico dos documentos nos processos de Recuperação de Informação. No entanto, os sistemas convencionais de extração de informação que estão baseados em palavras-chave não consideram que as palavras podem representar entidades geográficas espacialmente relacionadas com outras entidades nos documentos. Para resolver esse problema, é necessário viabilizar o georreferenciamento dos textos, ou seja, identificar as entidades geográficas presentes e associá-las com sua correta localização espacial. A identificação e desambiguação das entidades geográficas apresenta desafios importantes, principalmente do ponto de vista linguístico,Assim, a principal estratégia para superar os*

*problemas de ambiguidade, compreende a identificação de evidências que auxiliem na identificação e desambiguação das localidades nos textos. O presente trabalho propõe uma metodologia que permite identificar e determinar a cobertura espacial dos documentos, denominada SpatialCIM. A metodologia SpatialCIM tem o objetivo de organizar os processos de resolução de topônimos. Para isso, foram propostas e desenvolvidas as abordagens de (1) Desambiguação por Pontos e a (2) Desambiguação Textual e Estrutural. Essas abordagens, exploram duas técnicas diferentes de desambiguação de topônimos, as quais, geram e desambiguam os caminhos geográficos associados aos topônimos reconhecidos para cada documento. A partir dos resultados obtidos, foi possível demostrar que as técnicas de desambiguação melhoram a precisão e revocação na classificação espacial dos documentos. Demonstrou-se também o impacto positivo do uso de uma ferramenta linguística no processo de reconhecimento das entidades geográficas. Assim, foi demostrada a utilidade dos processos de desambiguação para a obtenção da cobertura espacial dos documentos.*

## 1. Introdução

A contínua evolução da tecnologia computacional tem possibilitado gerar e processar uma grande quantidade de dados. Como consequência, e aliada ao crescente aumento da capacidade de armazenamento, tem se tornado comum a utilização de grandes bases de dados em diversas áreas da atividade humana. Esses avanços tecnológicos têm causado um problema conhecido como superabundância de dados, pois nossas capacidades de coletar e armazenar dados têm superado à habilidade de analisar e extrair conhecimento dos mesmos [Fayyad et al. 1996, Cardoso 2011]. Nesse sentido, faz-se necessária a aplicação de técnicas e ferramentas que sejam capazes de extrair dos dados as informações úteis que representem conhecimento [Brachman and Anand 1996, Overell 2009]. Apesar de iniciativas recentes, os mecanismos de extração de informação apresentam deficiência na recuperação de conteúdos semânticos [Baeza-Yates et al. 2008], como informações geográficas relacionadas ao contexto [Jones and Purves 2008, Martins and Calado 2011].

O processo de identificação do contexto geográfico de textos é denominado *geotagging* [Amitay et al. 2004] e envolve duas etapas principais, *geo-parsing* e *geo-coding* [McCurley 2001]. A primeira, envolve a identificação das entidades geográficas presentes nos textos por meio da análise do seu conteúdo. Esses processos são complicados devido ao fato de que as palavras analisadas podem apresentar diversos tipos de ambiguidade. A segunda, tem o objetivo de identificar o contexto ou extensão geográfica do texto por meio de coordenadas geográficas.

A ambiguidade é uma expressão de linguagem que tem diferentes significados e pode ser entendida de diversas formas por um receptor. Esse problema causa ruido nos processos de recuperação de informação, já que o mesmo termo pode ter associado informação relevante ou irrelevante [Cardoso 2011]. Segundo [Silva 2006], a ambiguidade está relacionada com o contexto, sendo possível extrair diversos significados de uma frase ou palavra. A resolução de ambiguidades por meio do processamento da linguagem natural enfrenta vários problemas de difícil resolução como o conhecimento de contexto que não está explícito no texto analisado e os diversos tipos de ambiguidade presentes nos textos.

Para resolver esses problemas devem ser aplicados processos de desambiguação de entidades, os quais são responsáveis por encontrar a localização espacial no texto por meio de padronização das entidades numa representação estruturada [Roberts et al. 2010]. Segundo [Clough et al. 2004] os tipos de ambiguidade no contexto geográfico podem ser: (i) Ambiguidade da referencia (ARC), acontece quando determinada localidade pode ser referenciada por vários nomes diferentes; (ii) Ambiguidade do referente (ART), acontece quando um nome pode ser usado para referenciar outras localidades; e (iii) Ambiguidade da classe do referente (ACR), que acontece quando o nome pode ser usado para referenciar outros tipos de entidades. A área que estuda a identificação e desambiguação de topônimos é definida por [Leidner 2006] como Resolução de Topônimos.

Na literatura, a maioria das pesquisas segue uma série de passos para resolver a ambiguidade de topônimos. Esses passos estão baseados no reconhecimento de entidades, obtenção dos dados associados à entidade usando recursos externos, e, finalmente, o processo de desambiguação das entidades candidatas. No entanto, não é apresentada uma metodologia que indique os passos e etapas a serem seguidos antes de realizar os processos de desambiguação. Assim, neste trabalho é proposta a metodologia SpatialCIM (*Spatial Coverage Identification Methodology*) que proporciona uma série de etapas com o objetivo de organizar os processos de resolução de topônimos. As etapas propostas envolvem os processos de *geo-parsing* e *geo-coding*.

O artigo está organizado da seguinte forma: Na seção 2 é apresentada a metodologia SpatialCIM; na seção 3 são apresentados os experimentos e os resultados obtidos. Finalmente, na seção 4 são apresentadas as conclusões.

## 2. Metodologia SpatialCIM

A metodologia para determinar a cobertura espacial dos documentos denominada SpatialCIM é formada por uma série de passos que permitem identificar e localizar geograficamente um conjunto de documentos considerando a estrutura hierárquica do Brasil. A estrutura considerada segue a estrutura geo-política do Brasil: "Macro-Região, Região, Estado, Meso-Região, Micro-Região, Município, Usina e Categoria". Neste trabalho considera-se que o conjunto de documentos que serve como entrada da metodologia encontra-se na língua portuguesa e são do âmbito agrícola, mais especificamente notícias da cana-de-açúcar. Na Figura 1 é ilustrada a metodologia SpatialCIM que é composta por três etapas: (i) Pré-processamento, (ii) Expansão dos dados e (iii) Desambiguação.

Como observado na figura, os dados de entrada são um conjunto de documentos, os quais ao final do processo são convertidos a uma representação que contêm seus respetivos caminhos geográficos desambiguados. Na primeira etapa da metodologia os documentos são pré-processados de forma que seja realizado um reconhecimento das entidades geográficas. Uma vez que as entidades foram reconhecidas procede-se com a expansão dos caminhos geográficos de cada entidade, ou seja, identificar a estrutura hierárquica à qual pertence a entidade. Para realizar a expansão dos caminhos geográficos a metodologia SpatialCIM permite faze-lo de duas formas: (1) Sistema de Informação geográfico e (2) Ontologia geográfica. A seleção de uma dessas formas de expansão de dados determinará o processo de desambiguação a ser utilizado. Uma vez que os caminhos geográficos foram obtidos, analisam-se as entidades ambíguas e soluciona-se usando uma

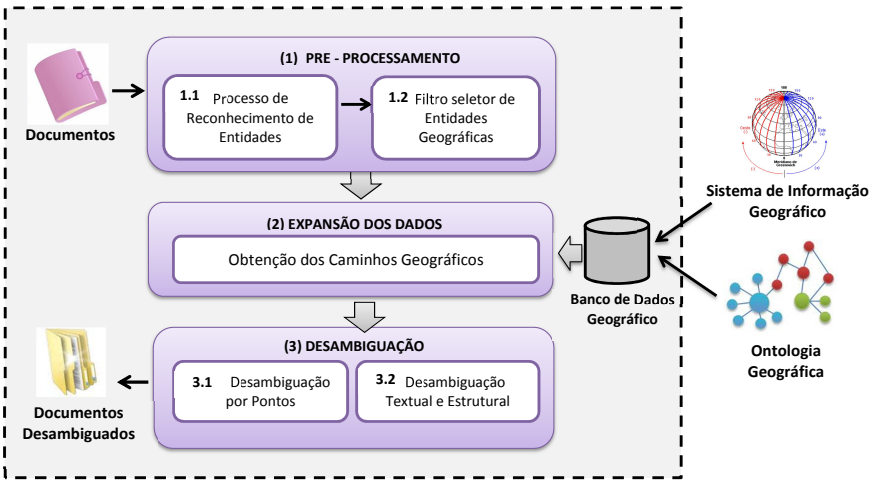


**Figura 1. Visão geral da metodologia SpatialCIM**

abordagem para desambiguar entidades.

## 2.1. Pré-processamento

Na etapa de pré-processamento o objetivo é analisar os documentos e conseguir extrair as entidades geográficas mencionadas. Esta etapa pode ser subdividida em (i) Processo de reconhecimento de entidades, que visa reconhecer todos os tipos de entidades presentes nos documentos, e (ii) Filtro seletor de entidades geográficas, que selecionará unicamente as entidades de carácter geográfico.

No reconhecimento de entidades, o objetivo é reconhecer as entidades presentes nos documentos para depois classificá-las em diferentes tipos. Esse reconhecimento de entidades pode ser realizado utilizando uma ferramenta linguística que analisa o contexto de cada palavra, ou fazendo uso de uma ferramenta baseada em regras que analisa as palavras para determinar se são ou não entidades geográficas. Devido ao fato que os documentos estão escritos em português a ferramenta linguística selecionada para efetuar o Reconhecimento de Entidades Mencionadas (REM) é o Rembrandt [Cardoso 2008]. Por outro lado, a ferramenta baseada em regras, é o Vocabulário Controlado AGRI-BR, o qual foi desenvolvido como parte deste trabalho. O Rembrandt permite o reconhecimento de diversas entidades, como números, organizações, pessoas e tempo. Depois de ter reconhecido as entidades mencionadas dentro dos documentos é necessário aplicar um (ii) filtro para extrair as entidades geográficas. O Filtro de Entidades Geográficas é aplicado unicamente quando os documentos foram processados pelo Rembrandt devido à variedade de entidades reconhecidas.

## 2.2. Expansão de Dados

Após a seleção das entidades geográficas é necessário obter as possíveis localidades em que a entidade pertence. Neste trabalho, as localidades associadas com as entidades são chamadas de caminhos geográficos. Para isso, é utilizado um banco de dados geográfico, mostrado na Figura 1. A principal diferença entre essas maneiras de desambiguar os textos é que o SIG usa como base para a desambiguação as coordenadas geográficas enquanto que a ontologia geográfica usa como base a estrutura hierárquica das entidades.

**Tabela 1. Exemplo dos caminhos geográficos extraídos para as entidades reconhecidas de um documento**

| | MACRO REGIÃO | REGIÃO | ESTADO | MESO REGIÃO | MICRO REGIÃO | MUNICIPIO | USINA | CATEGORIA |
|---|---|---|---|---|---|---|---|---|
| A1 | Nordeste | Nordeste | Alagoas | - | - | - | - | - |
| B1 | Nordeste | Norte | Pará | Metropol. Belém | Belém | Belém | - | - |
| B2 | Amazônia | Norte | Pará | Metropol. Belém | Belém | - | - | - |
| B3 | Amazônia | Nordeste | Paraíba | Agreste Paraibano | Guarabira | Belém | - | - |
| B4 | Nordeste | Nordeste | Alagoas | Agreste Alagoano | Palmeira dos Índios | Belém | - | - |
| C1 | Nordeste | Nordeste | Maranhão | Leste Maranh. | Chapadas do Itapecuru | Nova Iorque | - | - |
| C2 | - | - | - | - | - | - | - | Internacional (Nova Iorque) |
| D1 | Nordeste | Nordeste | Pernambuco | Mata Pernamb. | Mata Meridional Pernamb. | Marial | Una Export | - |
| E1 | Centro-Sul | Sudeste | São Paulo | Metropolit. São Paulo | São Paulo | São Paulo | - | - |
| E2 | Centro-Sul | Sudeste | São Paulo | Metropolit. São Paulo | São Paulo | - | | - |
| E3 | Centro-Sul | Sudeste | São Paulo | - | - | - | - | - |

Na Tabela 1 é mostrada a expansão dos caminhos geográficos realizadas para as entidades geográficas reconhecidas de um determinado documento. Observa-se que as entidades (b) Belém, (c) Nova Iorque e (e) São Paulo têm vários caminhos associados, porém são consideradas como entidades ambíguas. Na tabela, podem-se verificar diferentes tipos de ambiguidades com as quais os desambiguadores têm que lidar e resolver: (i)**Ambiguidade de nomes**, é o tipo de ambiguidade mais comum que acontece. Por exemplo, a entidade "Belém" é associada a diferentes caminhos geográficos com diferentes níveis de hierarquia, como "B1", "B2", "B3" e "B4"; (ii)**Ambiguidade hierárquica**, está representado na tabela pela entidade "São Paulo", na qual se tem os mesmos caminhos geográficos ("E1", "E2" e "E3"), mudando unicamente o nível da hierarquia reconhecido; e (iii)**Ambiguidade local-internacional**, está representada pela entidade "Nova Iorque", que pode pertencer tanto a um município do Brasil ou a uma entidade internacional. Dentro do Brasil, existem diversas entidades geográficas que compartem o nome com entidades internacionais tais como França, Califórnia, entre outras. Um fato representativo deste tipo de ambiguidade é a forma de descrever a entidade no texto. Por exemplo, se no texto a entidade tivesse aparecido como "New York" não teria sido reconhecida como ambígua.

## 2.3. Desambiguação

Uma vez que os caminhos geográficos foram obtidos usando alguma estratégia de extração de dados, é necessário aplicar um processo de desambiguação nas entidades que foram previamente detectadas como ambíguas. Na metodologia SpatialCIM são consideradas duas formas de efetuar a desambiguação: (1) Desambiguação por Pontos e (2) Desambiguação Textual e Estrutural.

### 2.3.1. Desambiguação por Pontos

O processo de (i) Desambiguação por Pontos é baseado no método proposto por [Leidner 2008]. Esse método usa como base as entidades não ambíguas reconhecidas e com base nelas desambiguam-se as outras entidades. Para efetuar a desambiguação é

necessário em primeiro lugar, obter as coordenadas geográficas de todas as entidades que têm caminhos geográficos associados.

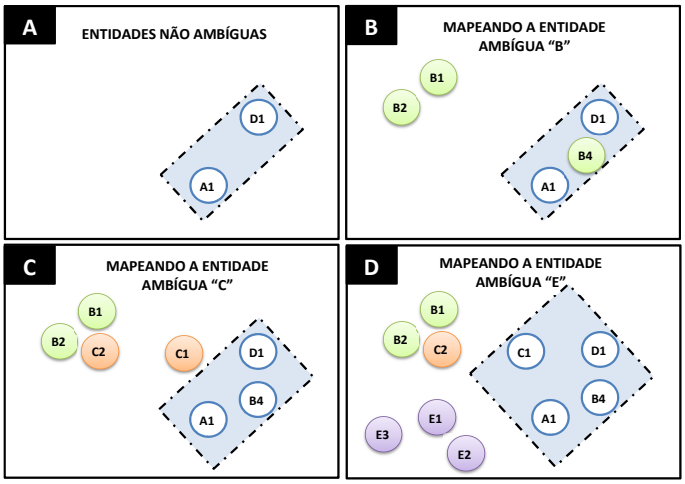


**Figura 2. Exemplo da construção de polígonos para o processo de Desambiguação por Pontos**

Para o processo de desambiguação das entidades apresentadas na Tabela 1, identificou-se que as entidades não ambíguas são (A1) Alagoas e (D1) Una Export. O primeiro passo, como ilustrado na Figura 2(a), é mapear no espaço o conjunto de entidades não ambíguas e construir um polígono entre elas. Após ter formado o polígono inicial, trabalha-se com cada conjunto de entidades ambíguas uma a uma. Na Figura 2(b) o processamento para a entidade "Belém" é mostrado. Esta entidade é representada na Tabela 1 pela letra "B". Os quatro caminhos geográficos dessa entidades são mapeados no mesmo espaço que foram mapeadas as entidades não ambíguas. Como observado, a entidade ambígua "B4" pertence ao polígono formado inicialmente pelas entidades não-ambíguas A1 e D1. Dessa forma, a entidade "B4" é reconhecida como a entidade desambiguada da entidade "B", por ser a entidade que pertence ao polígono das entidades não ambíguas. Após esse processo, um novo polígono é formado considerando agora as entidades "A1", "D1" e "B4". Na Figura 2(c) ilustra-se a construção do polígono para o conjunto de entidades ambíguas "C1" e "C2". Observa-se que nesse caso nenhuma das entidades pertence ao polígono formado no passo anterior. A entidade mais próxima de qualquer entidade que forma o polígono será considerada como a entidade desambiguada. No exemplo, a entidade "C1" é a entidade desambiguada. Com essa nova entidade, um novo polígono é formado e repete-se a operação até não encontrar mais entidades ambíguas, conforme ilustrado na Figura 2(d).

### 2.3.2. Desambiguação Textual e Estrutural

A desambiguação Textual e Estrutural não mapeia as entidades no espaço e, portanto não faz uso das coordenadas geográficas. Nesta abordagem de desambiguação faz-se primeira uma desambiguação usando o documento fonte e depois se utiliza a hierarquia das entidades para determinar a entidade desambiguada. Nos textos, é muitas vezes comum encontrar palavras ou termos associados que dão indicação de uma aproximação ou

especificação de um lugar geográfico. Segundo [Campelo and Baptista 2009], uma característica dos documentos que pode ser analisada é a ocorrência de Termos Especiais (ST - *Special Terms*). ST é que é definido como um termo cuja ocorrência em um texto pode ser decorrente da presença de referências geográficas no mesmo texto. Por exemplo, "cidade de Quirinópolis", "estado de Goiás" e "grupo Cosan". Os Termos Especiais aparentam fornecer uma melhor localização da entidade dentro do contexto do documento. Por isso, nesta abordagem de desambiguação, utiliza-se uma janela de tamanho 2. Essa janela serve para olhar se existem Termos Especiais antes ou depois da entidade geográfica. Escolheu-se uma janela de tamanho 2 já que usualmente os Termos Especiais estão o mais perto da palavra, por exemplo, "grupo Cosan", "estado de Alagoas".

O processo de desambiguação textual é realizado nas seguintes etapas: (i) Analisa cada documento com a finalidade de encontrar o lugar no qual está referenciada a entidade geográfica (linhas ou orações); (ii) Uma vez encontrada a entidade, procura-se 2 palavras para frente e 2 para trás com a finalidade de encontrar algum termo especial; e (iii) Para todos os Termos Especiais encontrados atribui-se um valor $\alpha$ que será denominado como "Peso Dica Texto". Esse valor serve para dar um maior peso a aquelas partes da hierarquia que foram referenciadas no texto. Por exemplo, assume-se que foi reconhecida a entidade "São Paulo" e têm associados dois caminhos geográficos. Um dos caminhos chega até o nível Estado e o outro até Município. Supondo que o termo especial "estado" foi encontrado, o caminho que chega até o nível estado tem mais probabilidade de ser o caminho correto. Assim, atribui-se o valor $\alpha$ para que, nesse contexto, "estado de São Paulo" tenha um maior peso que "município de São Paulo".

**CAMINHOS GEOGRÁFICOS**

| | MACRO REGIÃO | REGIÃO | ESTADO | MESO REGIÃO | MICRO REGIÃO | MUNICÍPIO | USINA | PAÍS |
|---|---|---|---|---|---|---|---|---|
| A1 | Nordeste | Nordeste | Alagoas | - | - | - | - | - |
| B1 | Nordeste | Norte | Pará | Metropolitana de Belém | Belém | Belém | - | - |
| B2 | Amazônia | Norte | Pará | Metropolitana de Belém | Belém | - | - | - |
| B3 | Amazônia | Nordeste | Paraíba | Agreste Paraibano | Guarabira | Belém | - | - |
| B4 | Nordeste | Nordeste | Alagoas | Agreste Alagoano | Palmeira dos Índios | Belém | - | - |

**PESOS HIERÁRQUICOS**

| | MACRO REGIÃO | REGIÃO | ESTADO | MESO REGIÃO | MICRO REGIÃO | MUNICÍPIO | USINA | PAÍS |
|---|---|---|---|---|---|---|---|---|
| A1 | β | β | β | - | - | - | - | - |
| B1 | β | 0 | 0 | 0 | 0 | α | - | - |
| B2 | 0 | 0 | 0 | 0 | 0 | - | - | - |
| B3 | 0 | β | 0 | 0 | 0 | α | - | - |
| B4 | β | β | β | 0 | 0 | α | - | - |

**Figura 3. Exemplo de Desambiguação estrutural das entidades**

Uma vez obtidos os Termos Especiais que se encontram no texto, procede-se com a desambiguação estrutural das entidades. Na Figura 3 é mostrado um extrato dos caminhos apresentados na Tabela 1. Pode-se observar que a desambiguação estrutural aplicada para a entidade ambígua "Belém" considera a entidade não-ambígua "Alagoas". A desambiguação estrutural pode ser dividida nos seguintes passos:

- Identificar o conjunto de entidades não ambíguas;
- Para cada parte da hierarquia das entidades não ambíguas atribui-se um valor $\beta$ que é denominado como "Peso Entidade não Ambíguo". Esse valor $\beta$ representa um peso para cada nível da hierarquia das entidades não-ambíguas e serve para

que as entidades ambíguas que compartem parte dos níveis não-ambíguos tenham um maior peso. Por exemplo, na Figura 3 para a entidade não-ambígua "Alagoas" atribui-se um valor $\beta$ para cada nível da hierarquia que faz parte seu caminho. Se observa-se as entidades ambíguas "B1", "B2", "B3" e "B4" nota-se que a entidade "B4" comparte os mesmos níveis que a entidade não-ambígua "Alagoas". Assim, atribui-se um valor $\beta$ para esses níveis. Nota-se também que a entidade "B1" comparte a mesma Macro-Região que a entidade de "Alagoas", porém, atribui-se um valor $\beta$ para a Macro-Região de "B1";

- Para todas as partes da hierarquia que tenham sido reconhecidas com Termos Especiais associa-se um valor $\alpha$, denominado "Peso Dica Texto". Por exemplo, assumindo que no texto encontrou-se o termo "município de Belém", todos os caminhos geográficos de "Belém" que cheguem até o nível de município terão associado um valor $\alpha$, como mostrado na Figura 3;
- Os níveis da hierarquia que não tenham um valor $\alpha$ ou $\beta$ associado terão um valor zero;
- Aplica-se, então, uma desambiguação heurística para cada conjunto de entidades. O objetivo dessa desambiguação é dar maior importância a entidades mais gerais. Assim, para a entidade "São Paulo" apresentada na Tabela 1 (entidade "E1", "E2" e "E3"), o melhor caminho geográfico é o apresentado pela entidade "E3". Essa desambiguação consiste em colocar valores de penalização à hierarquia; os valores considerados neste trabalho e considerando o domínio da cana-de-açúcar são: (1) Município: -9; (2) Micro-Região: -8; (3) Meso-Região: -7 e (4) Usina: +10. O nível de usina tem um valor positivo porque no escopo deste trabalho e do domínio dos documentos espera-se classificar os documentos considerando como parte importante as usinas reconhecidas. Esses valores de penalização foram escolhidos considerando que a desambiguação heurística tem um menor peso comparada com a desambiguação baseada no texto e na estrutura hierárquica das entidades. A finalidade desses valores é permitir variar um pouco a soma total para que caminhos genéricos tenham maiores valores e ao mesmo tempo não afete os valores $\alpha$ e $\beta$ atribuídos;
- Finalmente, todos os valores são somados e obtém-se um valor total para cada entidade. Escolhe-se aquela entidade que tem o maior valor e depois compara-se com um valor de descarte $\delta$. Se a entidade tiver um valor menor igual que $\delta$ é então descartada já que poderia se tratar de uma entidade que foi mal reconhecida.

Ao finalizar o processo de Desambiguação Textual e Estrutural, existe a possibilidade de ter eliminado várias entidades que podem ter sido bem ou mal reconhecidas nos processo de REM. Existe também a possibilidade de zerar todas as entidades reconhecidas se os valores das variáveis $\alpha$, $\beta$ e $\delta$ forem muito altos. Por isso, é importante ajustar os valores de $\alpha$, $\beta$ e $\delta$ até obter uma configuração apropriada para o problema, permitindo a eliminação de entidades que estejam abaixo de um limiar de eliminação determinado no início.

## 3. Experimentos e Resultados

Para a realização destes experimentos utilizou-se um banco de documentos provido pela Embrapa Informática Agropecuária[1] , que contém 698 notícias agrícolas da

[1] `http://www.cnptia.embrapa.br/`

cana-de-açúcar em português. Esse conjunto de documentos foi previamente marcado por um especialista da Embrapa, o qual serviu para validar os resultados obtidos pelos desambiguadores.

Na Tabela 2 é apresentado um resumo das medidas de precisão, cobertura e *F-score* para o conjunto de documentos comparando os processos não-desambiguado e desambiguado com a utilização da ferramenta Rembrandt e o AGRI-BR. Os resultados mostram um melhor desempenho dos documentos que foram pré-processados utilizando a ferramenta Rembrandt e passaram pelo processo desambiguação. Para os documentos pré-processados com o AGRI-BR, observou-se melhores resultados no processo desambiguado. O Rembrandt usando a Desambiguação por Pontos obteve um F-Score de 0.4862 e o AGRI-BR obteve 0.3836, demonstrando uma melhor performance dos documentos desambiguados e pré-processados com ajuda do Rembrandt.

**Tabela 2. Avaliação da Precisão, Cobertura e F-Score dos processos não-desambiguado e desambiguado para a Desambiguação por Pontos**

| | Rembrandt | | AGRI-BR | |
|---|---|---|---|---|
| | **Não-Desamb.** | **Desamb.** | **Não-Desamb.** | **Desamb.** |
| **Precisão** | 0.3147 | **0.4196** | 0.2131 | **0.2590** |
| **Cobertura** | **0.6702** | 0.5780 | **0.9047** | 0.7391 |
| **F-Score** | 0.4283 | **0.4862** | 0.3450 | **0.3836** |

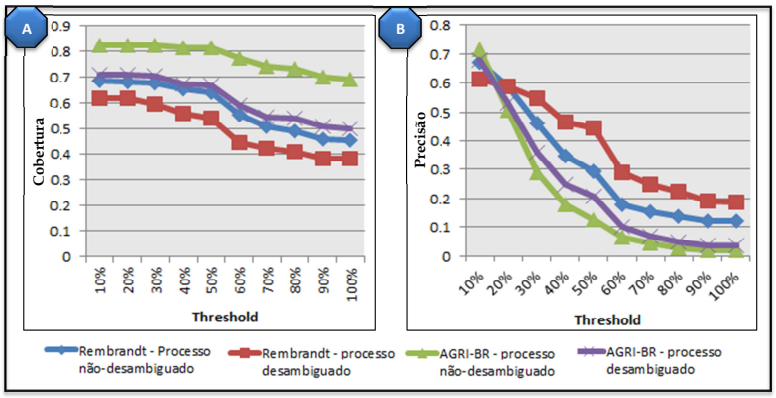


**Figura 4. Precisão e Cobertura para os processos não-desambiguados e desambiguados usando o Rembrandt e o AGRI-BR na Desambiguação por Pontos**

Na Figura 4 é ilustrado o comportamento do processo não-desambiguado e o processo desambiguado usando os dois tipos de pré-processamento definidos. Na Figura 4(a) é ilustrado o comportamento comparando a cobertura de todos os processos, e a Figura 4(b) ilustra o comportamento utilizando a precisão. Como observado na Figura 4(a), o processo que tem uma maior cobertura e recupera uma maior quantidade de entidades é o "AGRI-BR não-desambiguado". No entanto, esse mesmo processo Figura 4(b) representa aquele que tem a menor taxa de precisão. Observa-se também que os processos não-desambiguados representam aqueles com maior cobertura e menor precisão. Por outro lado, os processos de desambiguação mostram uma cobertura e uma precisão média.

O fato do *F-Score* aumentar para o Rembrandt e diminuir para o AGRI-BR devem-se à diferença entre a quantidade de entidades reconhecidas e a quantidade de entidades

corretamente reconhecidas por cada processo. Nesse caso, observa-se que os processos que foram pré-processados com o Rembrandt tendem a ter menor quantidade de entidades reconhecidas e maior quantidade de entidades corretamente reconhecidas.

**Tabela 3. Avaliação da Precisão, Cobertura e F-Score dos processos não-desambiguado e desambiguado para a Desambiguação Textual e Estrutural**

| | Rembrandt | | AGRI-BR | |
|---|---|---|---|---|
| | **Não-Desamb.** | **Desamb.** | **Não-Desamb.** | **Desamb.** |
| **Precisão** | 0.3157 | **0.4829** | 0.2052 | **0.2711** |
| **Cobertura** | **0.6965** | 0.5134 | **0.8594** | 0.5717 |
| **F-Score** | 0.4344 | **0.4977** | 0.3313 | **0.3678** |

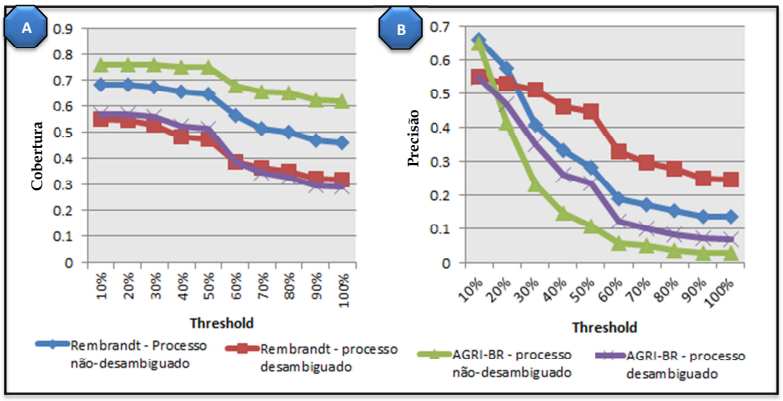


**Figura 5. Precisão e Cobertura para os processos não-desambiguados e desambiguados usando o Rembrandt e o AGRI-BR na Desambiguação Textual e Estrutural**

Na Tabela 3 é apresentado um resumo das medidas de precisão, cobertura e *F-score* aplicadas para o conjunto dos 698 documentos, comparando os processos não-desambiguados e desambiguados com a utilização da ferramenta Rembrandt e o AGRI-BR. Os resultados mostram um melhor desempenho dos documentos que foram pré-processados utilizando a ferramenta Rembrandt e passaram pelo processo desambiguação. No caso dos documentos pré-processados com o AGRI-BR, observa-se também melhores resultados no processo desambiguado. De forma que o Rembrandt usando a Desambiguação Textual e Estrutural obteve um *F-Score* de 0.4977 e o AGRI-BR obteve 0.3678. Nota-se que nesta abordagem melhoraram-se um pouco os valores de *F-Score* comparados com a Desambiguação por Pontos que obteve para o Rembrandt 0.4862 e para o AGRI-BR 0.3836.

Na Figura 5 é ilustrado o comportamento do processo não-desambiguado e o processo desambiguado usando os dois tipos de pré-processamento definidos. Na Figura 5(a) é ilustrado o comportamento dos processos de desambiguação e não-desambiguação comparando a cobertura. A Figura 5(b) ilustra o comportamento dos processos de desambiguação e não-desambiguação utilizando a precisão. Como pode ser observado na Figura 5(a) o processo que tem uma maior cobertura e recupera uma maior quantidade de entidades é o "AGRI-BR não-desambiguação" seguido pelo "Rembrandt não-desambiguado". Na Figura 5(b) observa-se que o "Rembrandt - desambiguado" apresentou melhores taxas de precisão que os outros algoritmos, seguido pelo "Rembrandt - não-desambiguado".

Nos dois tipos de desambiguação apresentadas, um fator importante que leva a cometer maior quantidade de erro está associado com o reconhecimento de entidades. Devido a que se uma entidade foi mal reconhecida os processos de desambiguação consideraram a entidade como parte do processo. Na Desambiguação por Pontos, nenhuma entidade é descartada. Assim, se uma entidade foi mal reconhecida o polígono gerado não representa a realidade dos documentos, desambiguando caminhos de forma errada. Na Desambiguação Textual e Estrutural utiliza-se o conceito de descartar entidades que encontram-se abaixo de um certo limiar. No entanto, esse limiar aplica-se unicamente às entidades ambíguas. Por tanto, se uma entidade mal reconhecida não é ambígua ela não será eliminada e será utilizada como base para a desambiguação.

## 4. Conclusoes

Neste artigo, foi apresentada a metodologia SpatialCIM que permite identificar e determinar a cobertura espacial dos documentos. A metodologia utiliza o *geo-parsing* e o *geo-coding* como parte das etapas. Dentro desta metodologia foram propostas duas formas de desambiguação que usa o conhecimento externo de *gazetteers* e ontologias geográficas, a Desambiguação por Pontos e a Desambiguação Textual e Estrutural. Como parte dos experimentos e da metodologia SpatialCIM, foi realizada uma comparação entre a ferramenta Rembrandt e o vocabulário controlado AGRI-BR para o processo de reconhecimento de entidades geográficas. Essa comparação foi realizada para avaliar se existe uma vantagem significativa no uso de uma ferramenta linguística sobre o uso de um vocabulário controlado. Para a avaliação experimental, foi utilizado um conjunto de documentos em Português que tratam sobre o tópico da cana-de-açúcar, e previamente marcado por um especialista da Embrapa. Finalmente, foram mostrados os resultados de cada abordagem de desambiguação, considerando as medidas de precisão, cobertura e F-Score. Adicionalmente foi apresentado o erro hierárquico para cada abordagem de desambiguação. Os experimentos mostraram um maior valor de F-score para os dados pré-processados com a ferramenta linguística Rembrandt. Foi mostrado também, que os processos de desambiguação melhoram a correta localização das entidades. Em trabalhos futuros, pretendem-se unir as duas abordagens de desambiguação, aproveitando a possibilidade de eliminar entidades da Desambiguação Textual e Estrutural e a vantagem de utilizar as coordenadas geográficas para determinar as localizações mais próximas da Desambiguação por Pontos.

## Agradecimentos

Queremos agradecer à Universidade de São Paulo (USP), ao Instituto de Ciências Matemáticas e de Computação (ICMC), Embrapa Informática Agropecuária, CNPQ e FAPESP pelo apoio.

## Referências

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 273–280.

Baeza-Yates, R. A., Ciaramita, M., Mika, P., and H., Z. (2008). Towards semantic search. *Proceedings of Natural Language and Information Systems*, 5039:4–11.

Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining*, pages 37–57.

Campelo, C. E. C. and Baptista, C. d. S. (2009). A model for geographic knowledge extraction on web documents. *Advances in Conceptual Modeling Challenging Perspectives*, 5833:317–326.

Cardoso, N. (2008). Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In *Encontro do Segundo HAREM (Avaliação de Reconhecedores de Entidades Mencionadas). In International Conference on Computational Processing of the Portuguese Language.*

Cardoso, N. (2011). Evaluating geographic information retrieval. *SIGSPATIAL Special*, 3:46–53.

Clough, P., Sanderson, M., and Joho, H. (2004). Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and R., U. (1996). From data mining to knowledge discovery: an overview. *Advances in knowledge discovery and data mining*, 1:1–37.

Jones, C. B. and Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, 22:219–228.

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30:400–417.

Leidner, J. L. (2008). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press.

Martins, B. and Calado, P. (2011). Learning to rank for geographic information retrieval. *Proceedings of the 6th Workshop on Geographic Informatino Retrieval GIR*, 1:1083–1084.

McCurley, K. (2001). Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 221–229. ACM.

Roberts, K., Bejan, C. A., and Harabagiu, S. M. (2010). Toponym disambiguation using events. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*. AAAI Press.

Silva, L. B. d. (2006). Ambiguidades da língua portuguesa: recorte classificatório para a elaboração de um modelo ontológico. Master's thesis, Universidade de Brasília - Departamento de Ciência da Informação e Documentação.