FILIPE CORRADINI

IMPLEMENTAÇÃO DE UM MÉTODO PARA EXTRAÇÃO DE CITAÇÕES BIBLIOGRÁFICAS USANDO UMA BASE DE CONHECIMENTO

LAVRAS - MG

2011

FILIPE CORRADINI

IMPLEMENTAÇÃO DE UM MÉTODO PARA EXTRAÇÃO DE CITAÇÕES BIBLIOGRÁFICAS USANDO UMA BASE DE CONHECIMENTO

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso para a obtenção do título de Bacharel em Ciência da Computação.

Orientador

Prof. Dr. Denilson Alves Pereira

LAVRAS - MG

2011

FILIPE CORRADINI

IMPLEMENTAÇÃO DE UM MÉTODO PARA EXTRAÇÃO DE CITAÇÕES BIBLIOGRÁFICAS USANDO UMA BASE DE CONHECIMENTO

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso para a obtenção do título de Bacharel em Ciência da Computação.

Aprovada em *27 de Junho de 2011*

Prof. Dr. Ahmed Ali Abdala Esmin

Prof. Msc. Tiago Amador Coelho

Prof. Dr. Denilson Alves Pereira

Orientador

LAVRAS - MG

2011

Aos meus pais Rodolfo e Adele.

AGRADECIMENTOS

Agradeço a meus pais, que nunca contiveram esforços para eu chegar até aqui, abdicando de sonhos pelo meu crescimento. Importantíssimos na minha vida, vão estar para sempre no meu coração.

Aos meus avós, minha irmã Fernanda e minha namorada Bruna.

Aos meus grandes amigos da República Galo Bravo: Alan (Lonttra), Gustavo (Bola), Matheus (Falamansa), Renato (Chico), Luiz (Tuxo), Fábio (Fabão), Caio (Copasa), Elder (Pinguela), Guilherme (Ardózzia), Thiago (KCT) e José Gustavo (Bridão).

Aos companheiros da turma 2007/2 e do Centro Acadêmico.

Aos Profs. Denilson, Ahmed e Tiago.

Aos amigos de Guariba/Lavras e a todos que não mencionei, mas que torceram ou ajudaram de alguma forma, muito obrigado.

RESUMO

Recuperação de Informação é uma área da computação que lida com recuperação automática de informações contidas em documentos. O principal objetivo deste trabalho foi elaborar e implementar um método para extrair citações de documentos de texto digitalizados, identificando cada citação contida neles. O método desenvolvido usa uma base de conhecimento com citações bibliográficas utilizadas para classificar cada *token* do documento lido e uma gramática livre de contexto que determina os padrões de referências para identificar as citações contidas nos textos. Após a elaboração e implementação, foi feita uma análise dos resultados obtidos com a execução do método desenvolvido. Foram feitos dois experimentos e calculados os valores de *Precision* (Precisão), *Recall* (Revocação) e *F-measure* sobre os dados de cada experimento, obtendo-se resultados satisfatórios.

Palavras-chave: Extração de Citação; Recuperação de Informação; Bibliotecas Digitais.

ABSTRACT

Information Retrieval is an area of computing that works with automatic retrieval of information contained in documents. The main objective of this study was to develop and implement a method to extract citation from scanned text documents, identifying each citation contained therein. The method uses a knowledge base of bibliographic citations used to classify each token of the document read and a context-free grammar that determines the patterns of references to identify the citations contained in the texts. After the development and implementation, was made an analysis of the results obtained with the execution of the method developed. It was made two experiments and calculated values of Precision, Recall and F-measure on the data of each experiment, getting satisfactory results.

Keywords: Extraction Citation; Information Retrieval; Digital Library.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 Contextualização

Recuperação de Informação é uma área da computação que consiste em recuperar automaticamente informações contidas em documentos de texto, imagem, som, vídeo etc.

As principais etapas da Recuperação de Informação são: busca pelos documentos de onde serão extraídas informações, busca pela informação desejada, extração destas informações e armazenamento das informações extraídas, se for o caso.

Essa área vem crescendo gradativamente devido à grande explosão no volume de documentos ao longo dos anos, principalmente na Web, o que explica a grande importância de serem desenvolvidas técnicas cada vez mais eficientes para atender as necessidades dos usuários.

Em todas as técnicas já desenvolvidas para extração de informações são encontradas várias informações indesejadas junto às extraídas, uma vez que existe mais de um padrão de escrita para o que se deseja extrair. Este é um grande problema que deve ser resolvido melhorando-se as soluções existentes ou até mesmo desenvolvendo novas técnicas.

Este problema é encontrado também no processo de extração de citações bibliográficas em documentos digitais, devido à existência de mais de um padrão para escrita de referencial bibliográfico. O termo citação é mais utilizado nessa área para se referir à referência bibliográfica.

A Tabela 1 ilustra exemplos de citações bibliográficas:

Tabela 1: Diferentes formas para escrita de citações.

BORKO, H. Information science: what is it? American Documentation, v.19, n.1, p. 3-5, 1968.
CASTI, J.L. Paradigms lost: images of man in the mirrar of science. New York: William Morrow, 1989.
DELIA, J.G. Communication research: a history. In: BERGER, C.R., CHAFFEE S.H. (Eds.) Handbook of communication science. Newbury Park, CA: Sage, 1978. P. 20-98.
Aha, D. and Kibler, D. (1991), Instance-based learning algorithms, Machine Learning, Vol. 6, pp. 37-66.
[Cardoso, O. 2003], Recuperação da Informação. Notas de Aula. UFLA, Brasil.
W. Goffman. Information science: discipline or disappearance. ASLIB Proceedings, v. 22 n.12, p. 589-596, 1970.

Com a Tabela 1, fica fácil perceber que cada citação possui informações não encontradas em outras, algumas têm volume, outras têm editora, outras têm páginas, etc.

Este trabalho propõe um método para identificar e extrair citações bibliográficas de documentos digitais tais como currículos, páginas de bibliotecas digitais e artigos científicos. O programa desenvolvido recebe como entrada um documento e retorna na saída as citações contidas nele, como mostra a Figura 1.

Uma vez que as informações a serem extraídas são buscadas em textos não estruturados, além dos diferentes padrões, é preciso levar em conta que uma citação pode ser encontrada em qualquer lugar do texto, o que dificulta ainda mais a identificação.

Quanto mais correta for essa extração, com maior precisão poderão ser captadas informações e estatísticas, como por exemplo, qual autor é o mais citado por outros, quantas publicações tem um determinado autor etc.

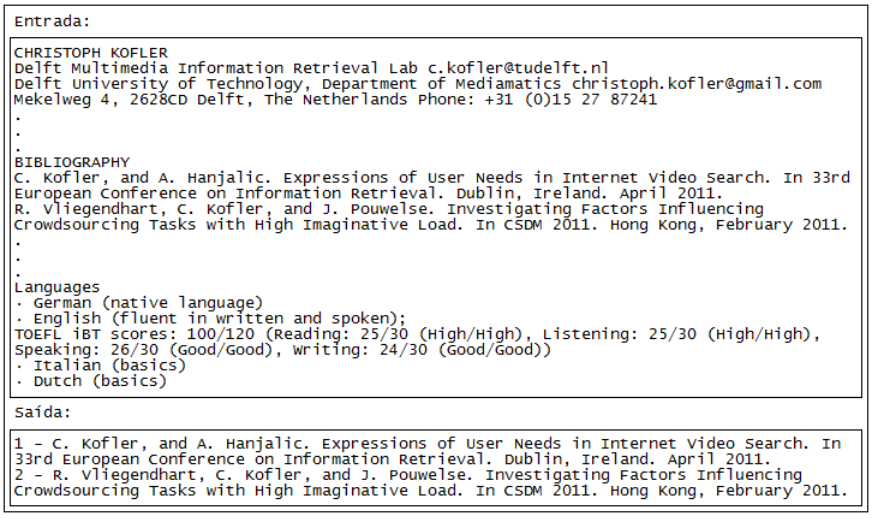

Figura 1: Exemplo de extração de citações de um currículo.

1.2 Objetivos Gerais e Específicos

O principal objetivo deste trabalho é elaborar e implementar um método para extrair citações de documentos de texto digitalizados, identificando cada citação contida neles.

Os objetivos específicos são: estudar métodos de extrações já existentes e fazer uma análise dos resultados obtidos com a execução do novo método desenvolvido.

1.3 Metodologia

Primeiramente foi feita uma revisão de bibliografia para o levantamento de algumas soluções já existentes para o problema descrito. Com base nesse levantamento, foi proposta uma solução específica para o problema de extração de

citações de currículos. A proposta implementada faz uma adaptação da estratégia usada pelo FLUX-CiM [3], que extrai atributos de citações, para a extração de citações de textos livres, tais como currículos de pesquisadores.

A solução proposta foi implementada e os resultados foram analisados a partir de dois experimentos. Primeiro foram selecionados os currículos de todos os professores do Departamento de Ciência da Computação da UFLA, que foram encontrados na Plataforma Lattes, e depois foram analisados curricula vitae, páginas da DBLP e artigos científicos ligados a área de recuperação de informação. Para esses experimentos, foram calculados os valores de *Precision* (Precisão), *Recall* (Revocação) e *F-measure*.

1.4 Organização Deste Documento

O restante deste documento está organizado da seguinte forma. O Capítulo 2 apresenta uma revisão da literatura sobre recuperação de informação, bibliotecas digitais e extração de dados de citações bibliográficas. O Capítulo 3 traz a descrição do trabalho com detalhes da implementação. No Capítulo 4 são apresentados os experimentos desenvolvidos e os resultados obtidos a partir deles. O Capítulo 5 apresenta as conclusões. E posteriormente encontram-se as referencias bibliográficas.

2 REFERENCIAL TEÓRICO

2.1 Recuperação de Informação

Recuperação de Informação é uma área da computação que lida com o armazenamento de documentos, seja este de texto, imagem, som, vídeo etc. e a recuperação automática de informação associada a eles. É uma ciência que estuda busca por informações em documentos, busca pelos documentos propriamente ditos, busca por metadados que descrevam documentos e busca em banco de dados, sejam eles relacionais e isolados ou banco de dados interligados em rede de hipermídia, tais como a *World Wide Web*. A mídia pode estar disponível sob forma de textos, de sons, de imagens ou de dados. Há, entretanto, muita confusão entre os termos e conceitos "recuperação de dados", "recuperação de documentos", "recuperação de informações" e "recuperação de textos". Na verdade, cada um destes é uma área especial que possui seu próprio corpo de conhecimento e literatura, teoria, praxis e tecnologias.

Segundo Ferneda [6], no contexto da Ciência da Informação, o termo "recuperação de informação" significa, para uns, a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário. Para outros, "recuperação de informação" consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes. O termo pode ainda ser empregado para designar a operação que fornece uma resposta mais ou menos elaborada a uma demanda, e essa resposta é convertida em um formato acordado com o usuário (bibliografia, nota de síntese, etc.). Há ainda

autores que conceituam a recuperação de informação de forma muito mais ampla, ao subordinar à mesma o tratamento da informação.

O processo de recuperação de informação consiste em identificar, no conjunto de documentos de um sistema, quais atendem à necessidade de informação do usuário. O usuário de um sistema de recuperação de informação está, portanto, interessado em recuperar "informação" sobre um determinado assunto e não em recuperar dados que satisfazem sua expressão de busca, nem tampouco documentos, embora seja nestes que a informação estará registrada. Essa característica é o que diferencia os sistemas de recuperação de informação dos sistemas gerenciadores de bancos de dados, estudados e implementados desde o nascimento da Ciência da Computação [6].

Os sistemas de banco de dados têm por objetivo a recuperação de todos os objetos ou itens que satisfazem precisamente às condições formuladas através de uma expressão de busca. Em um sistema de recuperação de informação essa precisão não é tão estrita. A principal razão para esta diferença está na natureza dos objetos tratados por estes dois tipos de sistema. Os sistemas de recuperação de informação lidam com objetos linguísticos e herdam toda a problemática inerente ao tratamento da linguagem natural. Já um sistema de banco de dados organiza itens de "informação", que tem uma estrutura e uma semântica bem definidas. Os sistemas de informação podem se aproximar do padrão que caracteriza os bancos de dados na medida em que sejam submetidos a rígidos controles, tais como vocabulários controlados, listas de autoridades, etc.

O termo Recuperação de Informação foi criado por Calvin Mooers em 1951 [11], e o campo de pesquisa é interdisciplinar, baseado em muitas áreas. Por

sua abrangência ele não é muito bem compreendido, sendo abordado tipicamente sob uma ou outra perspectiva. Ele está posicionado na junção de muitos campos já estabelecidos, tais como psicologia cognitiva, arquitetura da informação, projeto da informação, comportamento da informação humana, linguística, semiótica, ciência da informação, ciência da computação, biblioteconomia e estatística.

Sistemas (automatizados) de recuperação da informação foram originalmente usados para gerenciar a explosão da informação na literatura científica na segunda metade do século XX. Muitas universidades e bibliotecas públicas usam estes sistemas para prover acesso a livros, jornais, periódicos e outros documentos.

2.2 Bibliotecas Digitais

2.2.1 Conceitos Gerais

Biblioteca digital é a biblioteca constituída por documentos primários, que são digitalizados quer sob a forma material (CD-ROM, DVD, etc.), quer em linha através da Internet, permitindo o acesso à distância. Este conceito inclui também a ideia de organização composta por serviços e recursos cujo objetivo é selecionar, organizar e distribuir a informação, conservando a integridade dos documentos digitalizados.

Leiner [8], diz que uma biblioteca digital é a coleção de serviços e de objetos de informação, com organização, estrutura e apresentação que suportam o relacionamento dos utilizadores com os objetos de informação, disponíveis direta ou indiretamente via meio eletrônico/digital.

Existem autores propondo definições radicais, rejeitando a existência de conceitos das bibliotecas tradicionais no âmbito das bibliotecas digitais, ou ne-

gando de todo a existência destas últimas, ao defenderem que os pressupostos são radicalmente diferentes dos das tradicionais. No entanto uma ideia é clara: da definição do conceito "biblioteca digital" deverá relevar a ideia que, embora não sendo uma continuação das bibliotecas tradicionais, as bibliotecas digitais não podem negar a sua "emulação" por estas constituírem uma base instalada de grandes dimensões. Por outro lado, não se pode limitar a evolução tecnológica das bibliotecas digitais pelo apego a eventuais conceitos ultrapassados associados às bibliotecas tradicionais.

Uma biblioteca digital permite o acesso remoto através de um computador com ligação em rede e, ao mesmo tempo, a sua utilização simultânea por diversos utilizadores, onde estes podem encontrar em suporte digital os produtos e serviços característicos de uma biblioteca física. Através dela, é também possível utilizar de forma integrada diferentes suportes de registo de informação (texto, som, imagem).

As bibliotecas digitais eliminam as barreiras físicas e a distância, fatores que desde sempre limitaram o âmbito das bibliotecas físicas - biblioteca sem muros. Porém, estas bibliotecas sofrem de outros tipos de limitações, nomeadamente em nível da sua temática.

Assim, a Internet, meio por excelência de transmissão da informação neste contexto, comporta diferentes aspectos únicos, como sendo a capacidade de memória, a transportabilidade e a ubiquidade da informação.

2.2.2 Exemplos

Alguns exemplos de Bibliotecas Digitais a nível internacional são mostrados na Tabela 2:

Tabela 2: Exemplos de Bibliotecas Digitais.

Biblioteca	Site
The British Library	http://portico.bl.uk/
National Library of Australia	http://www.nla.gov.au/
Royal Library Belgium	http://www.kbr.be/
The Princess Grade Irish Library of Monaco	http://www.pgil.mc/
The WWW Virtual Library	http://vlib.org/
Library of Congress	http://lcweb.loc.gov/
Library and Archives Canada	http://www.nlc-bnc.ca/
Bibliothèque Nationale de France	http://www.bnf.fr/
Biblioteca Virtual de Macau	http://www.macaudata.com

2.2.3 Problemas de Ambiguidade com Citações

As informações presentes em bibliotecas digitais normalmente são obtidas por alguns métodos típicos. O auto arquivamento consiste na submissão de metadados e textos completos ao repositório pelos próprios pesquisadores. Silva [13] propõe um serviço de auto arquivamento de metadados para a Biblioteca Digital Brasileira de Computação (BDBComp).

Podemos também destacar esforços de diferentes comunidades para simplificar a aquisição de novas informações em bibliotecas digitais. A *Open Archives Initiative* (OAI - http://www.openarchives.org/), iniciada em outubro de 1999, for-

nece alguma assistência em relação a este problema. A abordagem OAI é baseada em, periodicamente, realizar a colheita (*harvesting*) de dados de diferentes fontes através de um protocolo simples e bem definido, denominado *Open Archives Initiative Protocol for Metadata Harvesting*. Os dados obtidos podem ser processados, integrados aos dados de outras origens e então carregados para o repositório da biblioteca digital. Considerando o modo como tais dados são obtidos, um problema imediato vem à tona. As fontes fornecedoras nem sempre mantêm compromisso com a padronização da escrita de seus dados. Além disso, duas fontes distintas podem utilizar padronizações diferentes. Por exemplo, enquanto uma fonte pode armazenar nomes de autores na forma "Último sobrenome, Primeiro nome, Iniciais dos nomes intermediários", outra fonte pode utilizar o padrão de escrita "Primeiro nome, Iniciais dos nomes intermediários, Último sobrenome" [10].

Ao realizarmos uma consulta em uma biblioteca digital, a consequência direta desse problema é a fragmentação das respostas obtidas. Entretanto, podemos verificar com pouco esforço que uma cadeia de caracteres se refere a diversas pessoas do mundo real. Nomes de autores sofrem problemas de variação causados por abreviações (José S. A. Silva ou José Silva), apelidos (William ou Bill), permutações (Bin Liu ou Liu Bin), grafias diferentes (Osvaldo, Oswaldo), uso de letras maiúsculas e minúsculas (José Da Silva, José da Silva), hiferação (Ribeiro-Neto, Ribeiro Neto), composição de nomes (El-Masri, Elmasri), uso de prefixos e sufixos (Sr., Jr., números, etc.), além de que algumas pessoas mudam de nome ao se casarem [9].

2.3 Extração de Dados de Citações Bibliográficas

2.3.1 Conceitos Gerais

Ao se trabalhar com dados em bases eletrônicas, pode-se distinguir formas em que é possível representá-las. Os dados estão dispostos na Internet e em outras fontes em três categorias, em relação à maneira como estão estruturados: semi-estruturados, estruturados e não-estruturados. As páginas da Web são consideradas dados "semi-estruturados", de caráter intermediário, ou seja, que apresentam "alguma estrutura". Têm-se ainda os dados "estruturados", como, por exemplo, aqueles presentes nos bancos de dados relacionais, e os "não-estruturados", como, por exemplo, o texto livre (currículos, artigos, etc.). Os documentos da Web apresentam variações, mas possuem alguma regularidade, ou seja, alguma estrutura. Têm sido estudados novos modelos de dados semi-estruturados e linguagens de consulta adaptadas a eles, visto que correspondem a uma representação mais flexível e mais adaptada ao ambiente da Web [1].

Muitas das páginas Web ricas em dados são geradas a partir de banco de dados estruturados, o que facilita o processo de extração dos dados nessas páginas. A despeito disso, é também comum que os dados presentes em páginas Web estejam dispostos em textos contínuos, onde não há delimitadores explícitos entre os dados, mas que escondem ainda assim uma estrutura implícita. Em [7], estes textos são chamados de textos semi-estruturados. Alguns exemplos deste tipo de texto são anúncios de classificados, endereços postais, referências bibliográficas, listas comerciais e currículos.

Extração de informação é o problema de obter a partir de documentos algumas informações específicas. Como por exemplo, obter citações bibliográficas

de um documento de texto qualquer. Neste caso, a parte do documento que não é relevante pode ser ignorada. Geralmente o problema é abordado no contexto de coleções específicas. Uma abordagem para o problema é percorrer todo o texto, encontrar palavras chaves e extrair a informação necessária dos contextos onde as palavras aparecem.

2.3.2 Exemplos

As Figuras 2 e 3 ilustram alguns exemplos de extrações em citações bibliográficas.

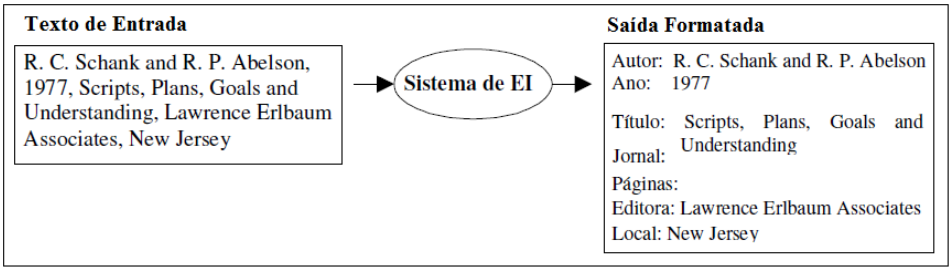

Figura 2: Exemplo de extração de informação em uma citação [12].

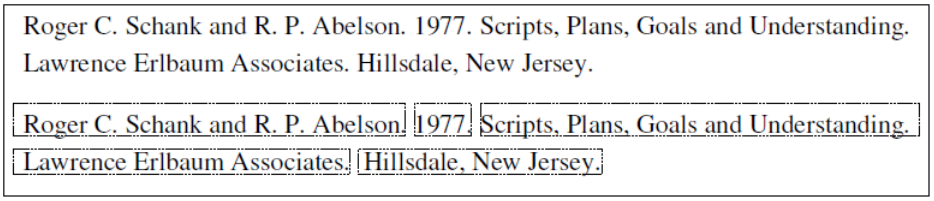

Figura 3: Exemplo simplificado de fragmentação de um texto de referência [12].

2.3.3 Abordagens Usadas por Alguns Autores

Em [12], é proposta uma abordagem híbrida para extração de informação que combina classificadores de texto e Modelos de Markov Ocultos (*Hidden Markov Models* - HMM). O classificador de texto gera uma saída inicial, que é refinada por meio de um HMM.

Em [4], é apresentada uma abordagem para extração em textos semiestruturados baseada em Modelos de Markov Ocultos (*Hidden Markov Models - HMM*). Essa abordagem dá ênfase à extração de metadados além dos dados propriamente ditos e consiste no uso de uma estrutura aninhada de HMMs, onde um HMM principal identifica os atributos no texto e HMMs internos identificam os dados e metadados (um pra cada atributo). Os HMMs são gerados a partir de um treinamento com uma fração de amostras da base a ser extraída. Um HMM é um autômato finito probabilístico, onde, os vértices são chamados de estados e as arestas são as transições entre os estados. Para cada aresta é associada uma probabilidade de transição. O autômato consome uma sequência finita de símbolos, ou observações, levando em consideração as probabilidades de transição de um estado para outro e as probabilidades de emissão, ou seja, a probabilidade de um determinado símbolo ser emitido por um estado específico.

Em [5], é criado manualmente um modelo que descreve um domínio de aplicação específico. Desse modelo, são identificados dados e palavras-chave no texto, através de regras de extração, ou expressões regulares, que são geradas. A identificação de palavras-chave é feita manualmente. Depois de identificadas, as palavras-chave são utilizadas exclusivamente para ajudar na extração de dados.

Em [2], é utilizada uma ferramenta de extração, denominada RAPIER, que gera expressões regulares a partir de treinamento supervisionado. Dados documentos-modelo, onde os dados de interesse são demarcados manualmente, esta ferramenta gera um conjunto de regras de extração através de um algoritmo baseado em lógica indutiva. As regras geradas pela ferramenta utilizam para extração as palavras e suas respectivas classes gramaticais. Para isso, é considerada uma estrutura gramatical fixa no texto.

Em [14], é utilizada uma gramática livre de contexto discriminativa para extrair o contato pessoal ou endereço de um texto semiestruturado, como documentos de texto e e-mails. Cada *token* do texto é rotulado de acordo com uma base de dados de treinamento, e seguindo a gramatica livre de contexto são procuradas sequências válidas, a fim de encontrar contatos pessoais e endereços a serem extraídos.

Em [3], é apresentado o FLUX-CiM, um novo método para extração de componentes de citações bibliográficas (nomes de autores, títulos de artigos, veículos de publicação, número de páginas, etc.). Este método não é limitado, ou seja, é possível extrair informações de citações em qualquer formato. O FLUX-CiM baseia-se em uma base de conhecimento automaticamente construída a partir de um conjunto existente de registros de metadados de um determinado campo (Ciência da Computação, Ciências da Saúde, Ciências Sociais, etc.). As etapas do FLUX-CiM são: etapa de blocagem, onde uma sequência de citações que contém os metadados a serem extraídos é dividida em unidades sintáticas chamadas de blocos; etapa de correspondência, onde os blocos são classificados de acordo com a base de conhecimento; etapa de ligação, onde blocos não classificados na

etapa anterior são analisados com base nas suas posições e as classificações dos "vizinhos"; etapa de adesão, onde os blocos são unidos formando o campo a ser extraído.

3 DESENVOLVIMENTO DO TRABALHO

3.1 Detalhes de Implementação

Este trabalho tem por objetivo desenvolver um método para extrair citações de documentos de texto digitalizados, identificando cada citação contida neles. O termo citação é o mais utilizado nessa área para se referir à referência bibliográfica. Para isso foi utilizada uma base de conhecimento e uma grámatica livre de contexto.

A base de conhecimento é uma base de dados que armazena termos que ocorrem em citações bibliográficas selecionadas. Estes termos são utilizadas como um "treinamento" para classificar cada *token* lido do documento. A idéia da gramática livre de contexto é de determinar os diferentes padrões de referências afim de identificar as citações contidas nos textos. A gramática identifica as citações analisando as classificações de cada termo do documento e a sequência em que se encontram.

O trabalho começa com uma fase de pré-processamento, onde é feito o preenchimento da base de conhecimento. Isto é feito via leitura dos dados. Neste caso foi lido um arquivo ".txt" com inúmeras referências bibliográficas no padrão bibtex, devido a sua padronização e ampla utilização no meio acadêmico. Porém não há impedimentos para que outros tipos de arquivos sejam utilizados.

A base de conhecimento foi implementada como uma tabela *hash* contendo termos que aparecem nas referências bibliográficas. Cada termo inserido na tabela *hash* possui uma lista de elementos que indicam quais atributos (autor, título, veículo de publicação, editora e endereço) em que o mesmo aparece, e o nú-

mero de ocorrências para tais atributos. Os artigos e preposições mais utilizados na língua portuguesa e inglesa (*stop words*) são desconsiderados, assim como os números e os símbolos.

Um exemplo do arquivo com as citações encontra-se na Figura 4.

```
@book{abadi1996theory,
  title = {A theory of objects},
  author = {Abadi, M. and Cardelli, L. },
  year = {1996},
  publisher = {Springer Verlag}
}
@article{aha.kibler.1991,
  title = {Instance-based learning algorithms},
  author = {Aha, D. W. and Kibler, D. and Albert, M. K. },
  journal = {Machine learning},
  volume = {6},
  number = {1},
  pages = {37--66},
  year = {1991},
  publisher = {Springer}
}
@article{alcantara.et.al.,
  title = {Recuperação de Informação},
  author = {d'Alcântara, A. A. and Silva, F. O. and Melo, L. P. and
Neto, M. and Souto, R. N. }
}
@book{allen1990probability,
  title = {Probability, statistics, and queueing theory: with computer
science applications},
  author = {Allen, A. O. },
  year = {1990},
  publisher = {Academic Pr}
}
@article{almeida.2002,
  title = {Uma introdução ao XML, sua utilização na Internet e alguns
conceitos complementares},
  author = {Almeida, M. B. },
  journal = {Ci. Inf},
  volume = {31},
  number = {2},
  pages = {5--13},
  year = {2002},
  publisher = {SciELO Brasil}
}
```

Figura 4: Modelo do arquivo bibtex.

Utilizando a primeira citação da Figura 4 como exemplo, os termos “theory” e “objects” seriam adicionados à base como título, os termos “Abadi”, “M”,

"Cardelli" e "L" seriam adicionados como autor, e "Springer" e "Verlag" como editora.

O trabalho de extração de citações bibliográficas começa com a leitura do arquivo de entrada (no formato texto) *token* por *token*, inserindo-os em ordem, em uma lista duplamente encadeada. Cada *token* é classificado como palavra (*token* do vocabulário português, inglês, etc.), número (*token* numérico), ou símbolo (*token* como ponto final, vírgula, dois pontos, etc.).

Cada *token* da lista duplamente encadeada do tipo palavra, é classificado de acordo com a base de conhecimento. Se o *token* é um termo que se encontra na base, verifica-se para qual atributo ele mais ocorreu, e classifica-o de acordo com este atributo. Com a base de conhecimento são classificados os atributos:"author" (autor), "title" (título da publicação), "venue" (veículo de publicação), "journal" (revista da publicação), "booktitle" (título dos anais da publicação), "publisher" (editora), "address" (endereço) e "location" (local).

Os atributos "year" (ano), "volume" (volume ou edição), "pages" (páginas) e "month" (mês), são classificados separadamente. São classificados como "year" os *tokens* numéricos de quatro algarismos entre 1900 a 2050, como "volume" os de um ou dois algarismos e como "pages" os de um a quatro algarismos com um traço entre os números. Todos os meses em português e inglês, bem como suas abreviações são classificados como "mouth".

As letras A, B, C, D, etc., seguidas de um ponto, são classificadas como autores, já que aparecem como abreviações de seus nomes. Portanto se o autor não existir na base de dados, provavelmente as abreviações existirão.

As Figuras 5 e 6, mostram a estrutura de dados da lista duplamente encadeada e da base de conhecimentos respectivamente.

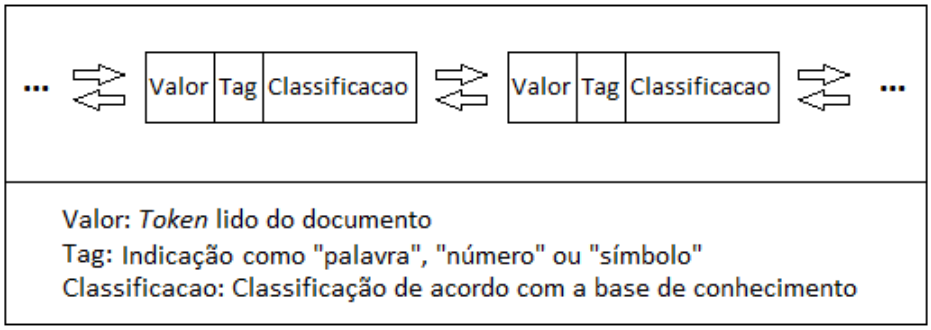

Figura 5: Estrutura de Dados da Lista Duplamente Encadeada.

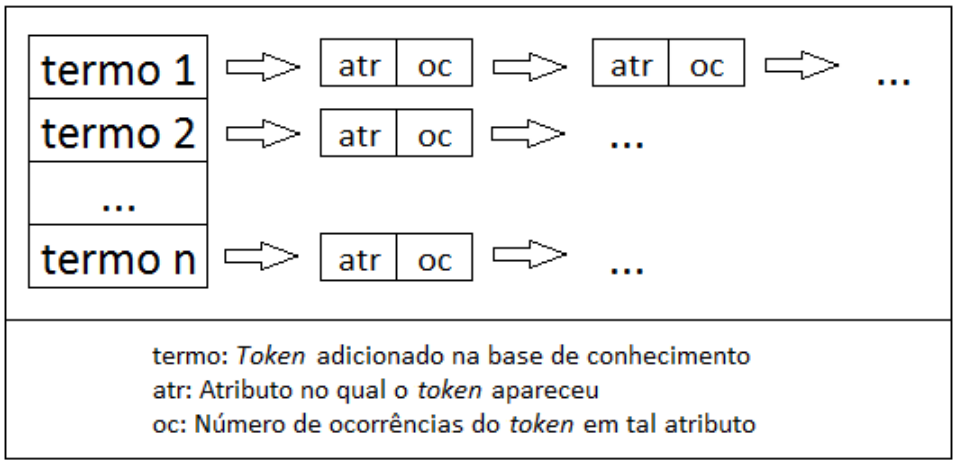

Figura 6: Estrutura de Dados da Tabela *Hash*.

Na Figura 5, o campo “Valor” de cada nó da lista duplamente encadeada indica o *token* lido do arquivo texto, o campo “Tag” indica a classificação deste *token* como “palavra”, “número” ou “símbolo”, e o campo “Classificacao” indica sua classificação de acordo com a base de conhecimento. Na Figura 6, os campos

“termo 1” até “termo n”, são os termos encontrados nas referências bibliográfias que alimentam a base de conhecimentos, e cada um destes termos apontam para uma lista de elementos que contém mais dois campos, “atr” e “oc”, que indicam respectivamente, o atributo no qual o termo apareceu, e o número de ocorrências do termo em tal atributo.

Em seguida, de acordo com uma gramática livre de contexto para citações bibliográficas, se existe uma sequência válida, uma citação é identificada.

A Figura 7 exemplifica uma gramática livre de contexto, onde uma referência bibliográfica pode tanto começar com autores quanto com título, seguido de veículo de publicação, volume, edição, páginas e ano, ou seguido apenas por veículo de publicação e ano, ou editora e ano, etc., mas sempre contendo, autor, título, veículo de publicação e ano.

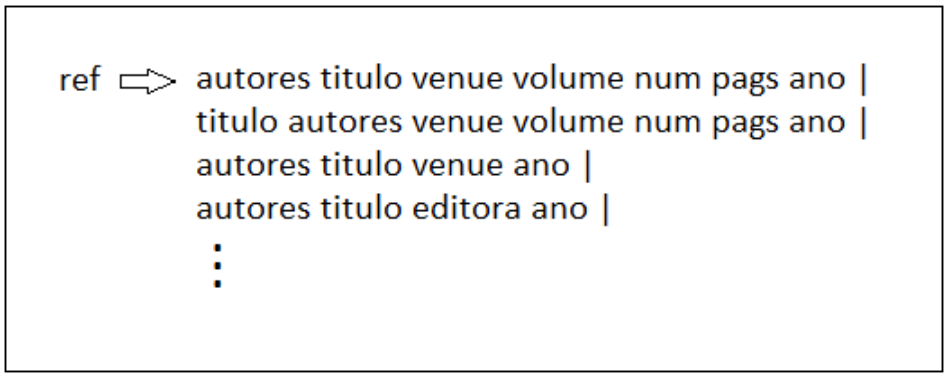

Figura 7: Exemplo de uma Gramática Livre de Contexto para Citações Bibliográficas.

É obrigatório a presença de alguma palavra classificada como autor, título, venue e ano, pois sem essa exigência, muitos itens irrelevantes são extraídos juntos aos resultados.

Depois de identificada uma citação, usando-se os símbolos separadores como auxiliares, e considerando também os tamanhos típicos de cada atributo, e um padrão para várias citações em sequência, são encontrados os limites (início e fim) da citação.

Quando uma citação é identificada e seu primeiro *token* é uma letra classificada como autor, assume-se que essa letra é a abreviação do nome de um autor que não está contido na base, então se o *token* anterior for uma "vírgula", o início da citação passa a ser o *token* anterior ao anterior.

Se uma citação é identificada e seus limites estão sendo analisados, se é encontrado um autor depois de venue, volume, etc., ou então um fim de linha seguido de um autor, determina-se o fim da citação analisada, e um possível início de outra em sequência.

3.2 Interface com o Usuário

O programa foi implementado no ambiente de desenvolvimento NetBeans e foi utilizada a linguagem Java.

A Figura 8 mostra a interface da saída do programa para o texto de entrada mostrado na Figura 1.

```
C:\Users\Filipe>java ExtracaoCitacao base.txt documento.txt
1 - c . kofler , and a . hanjalic . expressions of user needs in
internet video search . in 33rd european conference on informati
on retrieval . dublin , ireland . april 2011 .
2 - r . vliegendhart , c . kofler , and j . pouwelse . investiga
ting factors influencing crowdsourcing tasks with high imaginati
ve load . in csmd 2011 . hong kong , february 2011 .
```

Figura 8: Interface da saída do programa.

4 EXPERIMENTOS E RESULTADOS

4.1 Experimentos

Foram feitos dois experimentos para a análise dos resultados. O primeiro foi feito com os Currículos Lattes dos professores do Departamento de Ciência da Computação da UFLA, somando 22 no total. O segundo foi feito com curricula vitae, páginas da DBLP e artigos científicos da área de recuperação de informação, contendo 50 documentos de cada tipo citado, totalizando 150 documentos. Para efeito de comparação, foi analisado manualmente quantas citações deveriam ser encontradas em cada documento.

A Tabela 3 ilustra o número de citações reais contidas no primeiro experimento, em cada tipo de documento e no total do segundo.

Tabela 3: Número de citações reais contidas em cada tipo de documento.

Documento	**Número de citações**
Currículo Lattes	2240
Total Exp. 1	**2240**
Curricula Vitae	2390
Páginas DBLP	1975
Artigos	1340
Total Exp. 2	**5705**

No primeiro experimento foi utilizada uma base de dados com 800 citações bibliográficas no modelo bibtex para alimentar a base de conhecimentos. Estas citações foram coletadas manualmente e abrangem uma grande parte das áreas da Ciência da Computação. No segundo experimento, foi utilizada uma base

de dados com 400 citações também no modelo bibtex, coletadas manualmente no Google Acadêmico, porém concentrada apenas na área de recuperação de informação.

4.2 Resultados

Para cada execução, foi feita uma análise manual dos resultados obtidos quantificando as citações que encontravam-se corretas e quais estavam incorretas. A Tabela 4 mostra os resultados das extrações dos experimentos.

Tabela 4: Resultados das extrações dos experimentos.

Documento	Número de Citações	Extraído Corretamente	Extraído Incorretamente	Total Extraído
Currículo Lattes	2240	1329	84	1413
Total Exp. 1	**2240**	**1329**	**84**	**1413**
Curricula Vitae	2390	1685	115	1800
Páginas DBLP	1975	1650	0	1650
Artigos	1340	980	85	1065
Total Exp. 2	**5705**	**4315**	**200**	**4515**

A Figura 9 exemplifica um arquivo de entrada para o programa e sua respectiva saída. São extraídas duas informações, a primeira extração não é uma citação, portanto é considerada incorreta, e a segunda foi extraída corretamente, pois indica uma citação contida no texto de entrada.

Foram calculados os valores de *Precision* (Precisão), *Recall* (Revocação) e *F-measure* para o total de dados, e no segundo experimento foram também calculados os valores das mesmas métricas para cada tipo de documento analisado. As

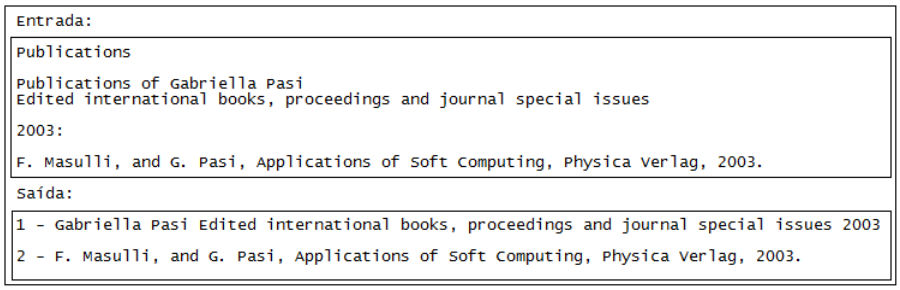

Figura 9: Exemplo de extração correta e incorreta.

fórmulas utilizadas para estes cálculos podem ser visualizadas na Figura 10, onde B é o cunjunto de citações que deveriam ser extraídas e S é o conjunto de itens extraídos, portanto a interseção de B e S indica os itens extraídos corretamente.

$$P = \frac{|B \cap S|}{|S|} \quad R = \frac{|B \cap S|}{|B|} \quad F = \frac{2(R.P)}{(R+P)}$$

Figura 10: Fórmulas de *Precision* (P), *Recall* (R) e *F-measure* (F).

A métrica *Precision* indica a porcentagem de acerto dos itens extraídos, a métrica *Recall* indica a porcentagem de citações encontradas, já *F-measure* é uma media harmônica ponderada entre *Precision* e *Recall*, seu valor varia entre 0 e 1, sendo melhor o resultado mais próximo de 1.

Após feitas todas as execuções, e feitos todos os cálculos, foram obtidos os resultados da Tabela 5.

Pode-se observar através da Tabela 5 que a taxa geral de Precisão do método desenvolvido e testado foi alto (variando entre aproximadamente 94 e 95%).

Tabela 5: Resultados das métricas dos experimentos.

Documento	***P***	***R***	***F***
Currículo Lattes	0,9405	0,5933	0,7276
Total Exp. 1	**0,9405**	**0,5933**	**0,7276**
Curricula Vitae	0,9361	0,7050	0,8043
Páginas DBLP	1	0,8354	0,9103
Artigos	0,9202	0,7313	0,8150
Total Exp. 2	**0,9557**	**0,7563**	**0,8444**

Já a taxa de Revocação deixou a desejar no primeiro experimento (aproximadamente 59%) e razoável no segundo (aproximadamente 84%).

No segundo experimento observa-se que a taxa de precisão para a extração nas páginas da DBLP é de 100%, isso acontece porque nestes documentos não contém outras informações além de publicações, o que torna o processo de extração mais simples.

A taxa de revocação aumentou no segundo experimento em relação ao primeiro, pelo fato da base de dados e os documentos do segundo estarem restritos apenas à área de recuperação de informação, enquanto os do primeiro abrangem mais áreas.

5 CONCLUSÃO

O presente trabalho teve como objetivo fazer um estudo das técnicas existentes para extração de informação em textos digitalizados e propor um método para extração de citações bibliográficas em documentos não estruturados. A principal motivação para esta proposta é que não se conhece nenhuma estratégia de se extrair citações de currículos. E isso é importante para a etapa seguinte, de extração dos atributos das citações, que é um trabalho já desenvolvido por alguns pesquisadores.

Com os resultados encontrados pode-se perceber que o método proposto e testado atingiu uma taxa de precisão de aproximadamente 94% no primeiro experimento e de aproximadamente 95% no segundo experimento. O que pode ser considerado satisfatório, pois minimizou as informações indesejadas junto às citações extraídas.

Já a taxa de revocação não se mostrou tão eficiente quanto a de precisão, uma vez que no primeiro experimento ficou em aproximadamente 59%, e no segundo experimento ficou em aproximadamente 76%, o que indica que várias citações não foram extraídas como deveriam. Esta diferença nos valores do primeiro experimento para o segundo, ocorre pelo motivo de o primeiro experimento abrangir uma grande parte das áreas da computação e o segundo estar restrito à apenas a área de recuperação de informação. Para haver uma melhora nos resultados deve-se aumentar o tamanho da base de conhecimentos, isso foi verificado fazendo testes conforme a base era aumentada, quanto maior era a base, mais citações eram extraídas.

Como trabalho futuro podem ser feitas melhorias ao método implementado, tentando maximizar a taxa de revocação sem interferir na taxa de precisão. Uma maneira de se fazer isso, além de aumentar o tamanho da base de conhecimentos, seria melhorar o tratamento da disposição dos símbolos separadores e das quebras de linha dentro de uma citação e uma melhor percepção do padrão de citações em sequência. Além disso, seria interessante fazer comparações dos resultados com outros métodos (*baselines*).

Referências

[1] ALMEIDA, M. Uma introdução ao xml, sua utilização na internet e alguns conceitos complementares. *Revista Ciência da Informação - Ci. Inf. 31*, 2 (2002), 5–13.

[2] CALIFF, M., AND MOONEY, R. Relational learning of pattern-match rules for information extraction. In *Proceedings of the 16th National Conference on Artificial Intelligence* (1999), JOHN WILEY and SONS LTD, pp. 328–334.

[3] CORTEZ, E., DA SILVA, A., GONÇALVES, M., MESQUITA, F., AND DE MOURA, E. A flexible approach for extracting metadata from bibliographic citations. *Journal of the American Society for Information Science and Technology 60*, 6 (2009), 1144–1158.

[4] DOS SANTOS, R., MESQUITA, F., DA SILVA, A., AND VILARINHO, E. Extração de dados e metadados em textos semi-estruturados usando HMMs. *Anais do XXI Simpósio Brasileiro de Bancos de Dados* (2006), 89–94.

[5] EMBLEY, D., CAMPBELL, D., JIANG, Y., LIDDLE, S., LONSDALE, D., NG, Y., AND SMITH, R. Conceptual-model-based data extraction from multiple-record web pages. *Data and Knowledge Engineering 31*, 3 (1999), 227–251.

[6] FERNEDA, E. *Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. PhD thesis, Universidade de São Paulo, 2003. Tese de doutorado.

[7] LAENDER, A., RIBEIRO-NETO, B., DA SILVA, A., AND TEIXEIRA, J. A brief survey of web data extraction tools. *ACM Sigmod Record 31*, 2 (Junho 2002), 84–93.

[8] LEINER, B. The scope of the digital library. *Draft for the DLib Working Group on Digital Library Metrics* (1998). http://www.dlib.org/metrics/public/papers/dig-lib-scope.html.

[9] LEY, M. The dblp computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval* (2002), vol. 2476, Springer, pp. 481–486.

[10] OLIVEIRA, J. Uma estratégia para remoção de ambiguidades na identificação de autoria de objetos bibliográficos., Abril 2005. Dissertação de Mestrado apresentada ao Programa de Pós Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais.

[11] SARACEVIC, T. Information science: Origin, evolution and relations. *International Conference on Conceptions of Library and Information Science: Historical, Empirical and Theoretical Perspectives. London: Taylor Graham* (1992), 5–27.

[12] SILVA, E., BARROS, F., AND PRUDÊNCIO, R. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. *XXV Congresso da Sociedade Brasileira de Computação* (Julho 2005), 504–513.

[13] SILVA, L. *Um Serviço de Auto-arquivamento de Publicações Científicas Compatível com o Padrão OAI*. PhD thesis, Universidade Federal de Minas Gerais, 2004. Dissertação de Mestrado, Departamento de Ciência da Computação, Belo Horizonte.

[14] VIOLA, P., AND NARASIMHAN, M. Learning to extract information from semi-structured text using a discriminative context free grammar. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (2005), ACM, pp. 330–337.