
Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

# Uma Abordagem Flexível para Extração de Metadados em Citações Bibliográficas

Eli Cortez Custodio Vilarinho

Manaus – Amazonas
Abril de 2009

Eli Cortez Custodio Vilarinho

# Uma Abordagem Flexível para Extração de Metadados em Citações Bibliográficas

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação e Banco de Dados.

Orientador: Prof. Altigran Soares da Silva, Doutor

Eli Cortez Custodio Vilarinho

# Uma Abordagem Flexível para Extração de Metadados em Citações Bibliográficas

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação e Banco de Dados.

Banca Examinadora

Prof. Dr. Altigran Soares da Silva – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Profa. Mirella Moura Moro, Ph.D.
Departamento de Ciência da Computação – UFMG

Prof. Dr. Edleno Silva de Moura
Departamento de Ciência da Computação – UFAM/PPGI

Prof. João Marcos Bastos Cavalcanti, Ph.D.
Departamento de Ciência da Computação – UFAM/PPGI

Manaus – Amazonas
Abril de 2009

*A Deus, a minha família pela dedicação de toda uma vida e a Lohaina, o amor da minha vida.*

# Agradecimentos

Inicialmente gostaria de agradecer a Deus por tudo que Ele tem feito em minha vida. Por todas as infinitas bênçãos e por todas as oportunidades. De forma alguma chegaria onde estou sem o Seu auxílio, sem a Sua salvação. Muito obrigado Senhor por me amar como eu sou e permitir que eu seja o Seu filho. Muito obrigado por minha família, Lohaina e meus amigos, pois através de cada um deles, entendo o quão imenso é o Teu cuidado e o Teu amor em minha vida. Eu te amo Senhor.

Tenho certeza de que palavras nunca serão suficientes para expressar a gratidão que eu tenho a Deus em relação a minha mãe, Raimunda Cortez, ao meu pai, Jone Reis, e a minha irmã, Gisele Cortez. Muito obrigado mãe, por todas os sacrifícios feitos, por todo o seu esforço, por todo o seu amor, carinho e dedicação. Muito obrigado por sempre acreditar em mim e nunca ter desistido. A semente do seu amor está plantada em minha vida e para sempre existirá em meu coração. Pai, muito obrigado por ser meu pai, ser você, literalmente você. Obrigado pelas sabatinas, ajuda com as lições, insistência, dedicação, amor. E Mana, muito obrigado por ser a minha irmã mais velha, por se preocupar comigo, por me amar insistentemente e incondicionalmente. Obrigado pelos incentivos, pelas conversas, pelas orientações. Minha família, onde estou ou chegarei, é o reflexo do amor, do cuidado e da dedicação de cada um de vocês. Muito Obrigado. Eu amo vocês.

Lohaina, obrigado por ter adicionado algo muito especial em minha vida, o amor. Nossa, o quão bom é poder ter a certeza de que se ama pra vida toda. Obrigado pelo incentivo, por acreditar em mim, e muito mais, por me permitir fazer parte da sua vida e me dar a possibilidade de compartilhar de seus sonhos para enfim, construirmos os nossos sonhos.

Obrigado pela paciência, carinho e pelas conversas. Eu amo você.

Professor Altigran, já se fazem quase 5 anos que trabalhamos juntos, e sem dúvida, o senhor tem grande participação em toda a minha carreira profissional. Obrigado por ter acreditado em mim desde quando ainda era um aluno da graduação. Obrigado pelos direcionamentos e por horas de dedicação. Obrigado por ter sido exigente pois assim entendi que o trabalho não vem do descanso. Saiba que o admiro bastante e tenho orgulho de ser um orientando seu. Enfim, obrigado!!!

Aos meus amigos, Guilherme, Mauro, Leonardo, Andrey, Rene, Raoni, Beto, Filipe, Wheidima, Michelle, David, Judson, André Garcia, André Carvalho, Klessius, Javier, Osman, Efren, Jonathas, Karane, Alessandro, Thomaz Philippe, muito obrigado por cada um de vocês que participam da minha vida. Obrigado Beto por ter sido o meu "Tutor" logo ao iniciar na pesquisa. Sem dúvida alguma, com você aprendi lições para toda vida, e além disso, ganhei um ótimo amigo. Obrigado Filipe, por na ausência do Beto, se fazer presente e se tornar também um guia, um amigo. Aos meus amigos da graduação resta dizer muito obrigado por tudo, por todas as experiências compartilhadas, por todos os momentos juntos, essencialmente, por tudo. Obrigado Mauro por estar sempre presente. Muito obrigado por ter sido um amigo e tanto. Obrigado Leonardo por ter mostrado que existem valores muito maiores do que simples habilidades profissionais. E ao Osman, muito obrigado por um ser grande amigo.

Aos meus amigos da igreja, obrigado em especial ao Judson, Wheidima e Michelle, por estarem presentes em minha vida, por serem meus amigos sempre e por me darem a certeza de que pra sempre estarão ao meu lado. E aos demais, saibam, que é um prazer, de alguma forma, fazer parte da vida de vocês.

Obrigado a todos!!!!

# Resumo

Nesta dissertação apresentamos o FLUX-CiM, um novo método de extração de componentes de citações bibliográficas tais como nomes de autores, títulos de artigo, números de página, etc. Tal método não se baseia em padrões específicos de codificação de delimitadores de um determinado estilo de citação, o que nos dá um alto grau de automação e flexibilidade e permite a extração de metadados a partir de citações em qualquer estilo. Diferentemente de abordagens anteriores que dependem de treinamento manual para realizar o reconhecimento de componentes em uma citação, no nosso caso, o método baseia-se em uma base de conhecimento automaticamente construída a partir de um conjunto existente de registros de metadados de um dado domínio, por exemplo: Ciência da Computação, Ciências da Saúde, Ciências Sociais, etc. Tal conjunto de registros com metadados pode ser facilmente obtido na Web ou através de outros repositórios de dados. Para demonstrar a eficácia e aplicabilidade do método proposto, apresentamos uma série de experimentos que visam extrair dados de citações bibliográficas de artigos. Os resultados destes experimento apresentam níveis precisão e revocação acima de 94% para todos os domínios, bem como extração perfeita para a grande maioria das citações testadas. Além disso, em uma comparação com o método que representa o estado da arte de extração de informação, o FLUX-CiM produziu resultados superiores sem a fase de treino que é exigida por esse método. Por fim, apresentamos uma estratégia para a utilização de dados bibliográficos resultante do processo de extração com FLUX-CIM para automaticamente atualizar e expandir a base de conhecimento de um determinado domínio. Mostramos que esta estratégia pode ser usada para alcançar bons resultados de extração mesmo quando apenas uma pequena amostra inicial de registros bibliográficos está disponível para a construção da base de conhecimento.

Palavras-chave: Gerenciamento de Citações, Extração de Metadados

# Abstract

In this dissertation, we present FLUX-CiM, a novel method for extracting components (e.g., author names, article titles, venues, page numbers) from bibliographic citations. Our method does not rely on patterns encoding specific delimiters used in a particular citation style. This feature yields a high degree of automation and flexibility and allows FLUX-CiM to extract from citations in any given format. Differently from previous methods that are based on models learned from user-driven training, our method relies on a knowledge-base automatically constructed from an existing set of sample metadata records from a given field (e.g., computer science, health sciences, social science, etc). These records are usually available on the Web or other public data repositories. To demonstrate the effectiveness and applicability of our proposed method we present a serie of experiments in which we apply it to extract bibliographic data from citations in articles of different fields. Results of these experiments exhibit precision and recall levels above 94% for all fields as well as perfect extraction for the large majority of citations tested. Also, in a comparison against a state-of-art information extraction method, ours produced superior results without the training phase required by that method. Finally, we present a strategy for using bibliographic data resulting from the extraction process with FLUX-CiM to automatically update and expand the knowledge-base of a given domain. We show that this strategy can be used to achieve good extraction results even if only a very small initial sample of bibliographic records is available for building the knowledge-base.

Keywords: Citation Management, Metadata Extraction

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

O gerenciamento de citações é um dos aspectos centrais nas bibliotecas digitais modernas. Citações [1] servem, por exemplo, como métrica para aferir do impacto ou da importância dos artigos científicos, e, portanto, da pesquisa que eles reportam. Avaliação do desempenho de indivíduos para promoções pode utilizar citações para medir a competência e o impacto do trabalho do pesquisador. Citações também têm sido utilizadas como fonte de evidências auxiliar em tarefas de Recuperação de Informação, tais como: classificação automática de documentos em [Calado et al., 2006, Couto et al., 2006], indexação e classificação [Lawrence et al., 1999] e avaliação da qualidade [Gonçalves et al., 2007]. Medidas Bibliográficas que baseiam-se em citações têm servido como inspiração para modernos algoritmos de análise de apontadores da Web, como *PageRank* apresentado em [Brin and Page, 1998]. Em um sentido mais amplo, as citações são a base de importantes projetos como: Digital Bibliography & Library Project (DBLP) [2] e Computer Science Bibliography[3].

O gerenciamento de citações em uma biblioteca digital envolve aspectos, tais como: (i) limpeza nos dados para correção de erros, atribuição imprópria de autoria ou divisão da produção um pesquisador devido à utilização de múltiplos nomes em publicações,

[1] Aqui interpretado como um conjunto de informações bibliográficas, tais como o nome do autor, título, local de publicação ou ano que são pertinentes a um artigo específico

[2] `http://www.informatik.uni-trier.de/~ley/db`

[3] `http://liinwww.ira.uka.de/bibliography`

e (ii) remoção de registros duplicados, principalmente após a integração de dados, ou a realização de entrada de dados. A maioria das técnicas que realizam essas tarefas baseia-se na suposição de que é possível identificar corretamente os principais componentes dentro de uma citação, como nomes de autores, título, local de publicação, ano, páginas, etc. Porém, esta não é uma tarefa simples por inúmeras razões como as identificadas por [Lee et al., 2007]: erros na entrada de dados, vários formatos de citação, a falta de uma norma, nomes de autores comuns, abreviaturas e grande volume de dados bibliográficos.

Nesta dissertação, apresentamos o método *FLUX-CiM (**Fl**exible **U**nsupervised E**x**traction – **Ci**tation **M**etada)*, um método para ajudar a extrair corretamente os componentes de citações bibliográficas. Diferentemente de abordagens anteriores apresentadas em [Embley et al., 1999, Day et al., 2005, Peng and McCallum, 2006], as quais dependem de treinamento manual para realizar o reconhecimento de componentes em uma citação, o nosso método baseia-se em uma base de conhecimento automaticamente construída a partir de um conjunto existente de registros de metadados de um dado domínio. Tal conjunto de registros pode ser facilmente obtido nos dias de hoje. Por exemplo, pode ser coletado diretamente a partir da Web ou a partir de repositórios abertos [OAI., 2005].

De forma geral, o nosso método de extração é baseado em: estimar a probabilidade de determinado termo encontrado em uma citação ocorrer como um valor de um dado campo bibliográfico de acordo com as informações encontradas na base de conhecimento, e usar propriedades estruturais genéricas das citações bibliográficas, tais como, o uso de sinais de pontuação para delimitar campos. Isto significa que a nossa abordagem não se baseia em padrões específicos de codificação dos delimitadores de um determinado estilo de citação, dando ao nosso método um elevado grau de automação e flexibilidade, como demonstrado através de experimentos.

Relatamos os resultados de experimentos com o nosso método para extrair informações de citações em três domínios diferentes. No domínio de Ciência da Computação usamos dados da coleção CORA [Peng and McCallum, 2006]. Em Ciências da Saúde, dados de vários artigos publicados pelo *National Institutes of Health (NIH)* foram utilizados, e

em Ciências Sociais utilizamos dados de vários artigos disponíveis na Biblioteca Digital Scielo. No caso da Ciência da Computação, para construir cada base de conhecimento, utilizamos os dados da própria coleção CORA. Em Ciências da Saúde e Ciências Sociais, utilizamos registros de metadados da *PubMed Central (PMC)* e Biblioteca Digital Scielo, respectivamente, ambos sendo repositórios digitais livres. Os resultados destes experimentos indicam que o método FLUX-CiM foi capaz de extrair corretamente, em média, mais de 94% dos valores dos campos presentes nas citações. Além disso, para mais de 82% das citações a extração foi perfeita, com todos os campos corretamente extraídos.

Também relatamos experimentos realizados para comparar nosso método, FLUX-CiM, com o CRF [Peng and McCallum, 2006], o estado da arte em extração de informação. Estes resultados corroboram as nossas afirmações em relação à elevada qualidade que o nosso método alcança, mesmo sem utilização de um treino manual. Em particular, FLUX-CIM obtém desempenho muito superior quando o conjunto de teste possui citações formatadas com vários estilos diferentes.

Por fim, apresentamos uma estratégia para a utilização de dados bibliográficos resultantes do processo de extração com FLUX-CIM para automaticamente atualizar e expandir a base de conhecimento de um determinado domínio. Mostramos que essa estratégia de realimentação pode ser utilizada para alcançar bons resultados mesmo se apenas uma amostra com poucos registros bibliográficos estiver disponível para a construção da base de conhecimento inicial.

O método aqui descrito foi publicado em três artigos [Cortez et al., 2007] [Cortez et al., 2009, Cortez and da Silva, 2008], os quais derivam deste trabalho de mestrado.

## 1.1 Trabalhos Relacionados

Nos últimos anos, várias ferramentas, métodos e técnicas têm sido propostos para solucionar a questão da extração de dados a partir de documentos textuais com enfoque nos documentos disponíveis na Web. Um breve estudo sobre este tema é apresentado

em [Laender et al., 2002b]. Para lidar com esse problema, várias técnicas distintas foram desenvolvidas, tais como: análise estrutural em documentos HTML [Crescenzi et al., 2001, Arasu and Garcia-Molina, 2003, Reis et al., 2004, Liu et al., 2003], processamento de linguagem natural [Freitag and McCallum, 2000, Muslea et al., 2001, Soderland, 1999], aprendizado de máquina [Hsu and Dung, 1998, Kushmerick, 2000], modelagem de dados [Laender et al., 2002a] e ontologias [Embley et al., 1999].

A maioria das abordagens anteriores utiliza fontes de treino com documentos (por exemplo, páginas da Web) rotulados com valores de exemplo, a partir das quais as regularidades na formatação em torno dos valores são aprendidas. Nessas abordagens, o processo de extração consiste em reconhecer e extrair *strings* dentro deste entorno que ocorrem nos documentos de entrada que são semelhantes aos documentos da fase de treinamento. Diferentemente, o método FLUX-CIM não depende de recursos de formatação de documentos de entrada (por exemplo, regularidades nos arredores dos valores ou na estrutura da página), mas sim no seu conteúdo, considerando as características dos campos bibliográficos juntamente com os seus valores. Assim, FLUX-CIM é capaz de reconhecer os valores para os campos bibliográficos, independentemente do formato específico dos documentos de entrada ou do estilo utilizado nos registros de citações.

Outra abordagem baseada em conteúdo para extração de dados foi proposta por [Embley et al., 1999]. Tal abordagem realiza a extração de dados baseada em ontologias utilizando um modelo semântico de dados para a construção de uma ontologia que descreve os dados de interesse, incluindo relacionamentos, aparência léxica e contextos de palavras-chave. Ao analisar esta ontologia, um esquema de banco de dados relacional e um reconhecedor de constantes/palavras-chave são gerados automaticamente, para então serem utilizados para extrair os dados que irão povoar o banco de dados. Esta abordagem baseia-se principalmente no conteúdo das páginas de acordo com aquilo que foi antecipado por uma ontologia pré-especificada construída por um especialista enquanto a maioria das abordagens utiliza o contexto textual em torno dos dados de interesse. Se a ontologia é suficientemente representativa, o processo de extração é totalmente automatizado. Neste

caso, o processo de extração é inerentemente resiliente e adaptável. Resiliente porque funciona corretamente mesmo se as características da fonte de formatação de documentos se alterarem. Adaptável porque funciona para os documentos de várias fontes distintas pertencentes a um mesmo domínio.

Na área de Bibliotecas Digitais, a extração automática de metadados é um campo que tem ganhado muita atenção recentemente. [Han et al., 2003] descrevem um método de extração de metadados em cabeçalhos de artigos científicos baseado no método de classificação de Support Vector Machines (SVM) e supera outros métodos de aprendizagem automática na mesma tarefa. MetaExtract é um sistema automático para atribuição de metadados que realiza a extração através de técnicas de processamento da linguagem natural aplicada à documentos de educação [Yilmazel et al., 2004]. Em [Hu et al., 2005], os autores focam na extração de títulos de documentos em geral (por exemplo, apresentações, capítulos de livro , trabalhos técnicos, brochuras, relatórios, cartas). Paynter [Paynter, 2005] aborda a avaliação automática de ferramentas que extraem metadados e discute suas vantagens e limitações. Em [Day et al., 2005] é proposta uma abordagem para extração de metadados baseada em um conhecimento ontológico com representação através de um *framework* chamado INFOMAP. Esta abordagem, assim como a proposta em [Embley et al., 1999], exige uma ontologia a ser construída. Neste caso, com a ajuda da ferramenta de edição Compass. Os autores relatam bons resultados de extração considerando 6 diferentes padrões fixo de citação somente para artigos de periódico.

Em [Peng and McCallum, 2006], os autores abordam o problema da extração de informações bibliográficas de artigos científicos e propõem a utilização de *Conditional Random Fields (CRFs)* para resolver este problema. CRF [Lafferty et al., 2001], é um modelo probabilístico comumente utilizado para extrair informações disponíveis em fontes textuais. Tal modelo funciona através da atribuição de rótulos a segmentos em um texto dado como entrada. As fases de rotulagem e de segmentação são baseadas em um modelo gerado a partir de um processo de treinamento onde instâncias do texto manualmente rotuladas e segmentadas. O treinamento tem como objetivo capturar vários aspectos

locais (por exemplo: seqüência dos campos, estilo de escrita), recursos léxicos externos (dicionários) layout e funcionalidades (por exemplo: pontuação, tipo de letra) para serem representados no modelo. Para demonstrar a qualidade do método proposto, os autores realizam experimentos com os dados da coleção CORA, que também foi utilizada em nossos experimentos. Atualmente, modelos baseados em CRF constituem o estado da arte em extração de informação devido à sua flexibilidade e qualidade dos resultados alcançados na extração. Assim, na seção de experimentos, apresentamos um estudo comparativo entre este método e a nossa abordagem.

A idéia da utilização de realimentação em extração de informação, embora não seja nova, apenas recentemente tem sido explorada na literatura. Foi anteriormente desenvolvida com CRF em [Culotta et al., 2006], onde os autores apresentam um estudo sobre como melhorar os modelos de extração utilizando realimentação de usuários. Este trabalho descreve um *framework* destinado a ajudar os usuários na correção manual dos modelos de extração baseados em CRF. Nesta dissertação, mostramos que o nosso método de extração (FLUX-CIM), permite a utilização de realimentação visando a melhoria na qualidade da extração através de maneira totalmente automatizada, ou seja, sem intervenção do usuário. Isto é uma importante distinção entre o trabalho apresentado em [Culotta et al., 2006] que exige um usuário para orientar manualmente o processo de realimentação.

## 1.2 Organização da Dissertação

Este trabalho está organizado da seguinte forma: O Capítulo 2 introduz os conceitos utilizados na nossa abordagem, apresenta o método proposto em detalhes e discute uma estratégia para automaticamente atualizar e expandir a base de conhecimento utilizando processo de realimentação. O Capítulo 3 detalha os experimentos, apresenta um estudo comparativo com o método estado da arte para extração de informação e apresenta resultados obtidos pelo processo de realimentação proposto. Por fim, o Capítulo 4 conclui a dissertação, indicando orientações para trabalhos futuros.

# Capítulo 2

# O Método FLUX-CiM

Neste capítulo, apresentamos os detalhes do nosso método de extração de metadados bibliográficos, FLUX-CiM, incluindo uma estratégia baseada em realimentação para atualizar e expandir a base de conhecimento utilizada pelo método. Inicialmente, introduzimos alguns conceitos e definições que serão utilizados durante toda esta dissertação. Em seguida, discutimos cada etapa[1] que compreende o nosso método. Primeiro, discute-se a etapa denominada *Blocking*, onde uma citação contendo os metadados a serem extraídos é dividida em unidades chamadas de *blocos*. Após a etapa de blocking, apresentamos a fase de *Matching*, que tenta associar os campos de metadados bibliográficos a cada bloco da citação através da informação disponível na base de conhecimento. Depois disto, discute-se a etapa de *Binding*, onde os blocos que não foram associados a nenhum campo bibliográfico na fase anterior são analisados e associados com base em sua posição relativa na citação. Então, discutimos a fase *Joining*. Nesta última etapa, os blocos são unidos para formar valores dos campos que compõem um registro de metadados. Após o detalhamento do processo de extração proposto, detalhamos nossa abordagem para atualização automática da base de conhecimento.

[1]Considerando que este método já foi anteriormente apresentado em artigos publicados em língua inglesa, resolvemos nesta dissertação manter os nomes originais em inglês utilizados para designar as fases do método.

## 2.1 Conceitos Básicos

### 2.1.1 Base de Conhecimento

Uma base de conhecimento é um conjunto de pares $BC = \{\langle m_1, O_1\rangle, \ldots, \langle m_n, O_n\rangle\}$, onde cada $m_i$ é um campo de metadados bibliográficos distinto, e $O_i$ é um conjunto de termos $\{o_{i,1}, \ldots, o_{i,n_i}\}$ chamado *ocorrências*. Intuitivamente, $O_i$ é o conjunto de valores típicos do campo $m_i$.

O processo de construção de uma base de conhecimento é trivial. Dado um conjunto de registros com metadados bibliográficos de uma determinada área, simplesmente processa-se cada registro e, para cada campo, extraímos os valores das ocorrências. Tal processo não requer nenhum esforço humano para selecionar os melhores registros que devem representar um dado domínio. Na verdade, é mais provável que o processo seja realizado automaticamente, usando conversores de formatos. Por exemplo, a base de conhecimento que construímos para nosso método para testar citações em Ciência da Computação veio de um conjunto de arquivos *bibtex* disponíveis na coleção CORA [Peng and McCallum, 2006]. Para isso, bastou-se analisar cada entrada e guardar os valores de cada campo na nossa base de conhecimento.

$$
\begin{aligned}
BC &= \{ \quad \langle Autor, O_{Autor}\rangle, \langle Titulo, O_{Titulo}\rangle\} \\
O_{Autor} &= \{ \quad \textbf{“J. K. Rowling”}, \textbf{“Galadriel Waters”}, \ \textbf{“Beatrix Potter”}\} \\
O_{Titulo} &= \{ \quad \textbf{“Harry Potter and the Half-Blood Prince”}, \\
&\qquad \textbf{“A Guide to Harry Potter”}, \textbf{“Petter Rabbit’s Halloween”}\}
\end{aligned}
$$

Figura 2.1: Exemplo de uma Base de Conhecimento.

Na Figura 2.1 é apresentado um exemplo simples de uma base de conhecimento, onde estão ilustrados somente dois campos de metadados bibliográficos: *Autor* e *Título*.

### 2.1.2 Citação

Uma *citação* é uma porção de texto englobando campos bibliográficos que pode ser obtida a partir de uma lista de referência bibliográficas em um artigo. Em nosso método, para a

obtenção de cada citação utilizamos simples conversores de formato que extraem o texto de arquivos em PDF e outros formatos populares. Na Figura 2.2(a) apresentamos um exemplo de uma citação.

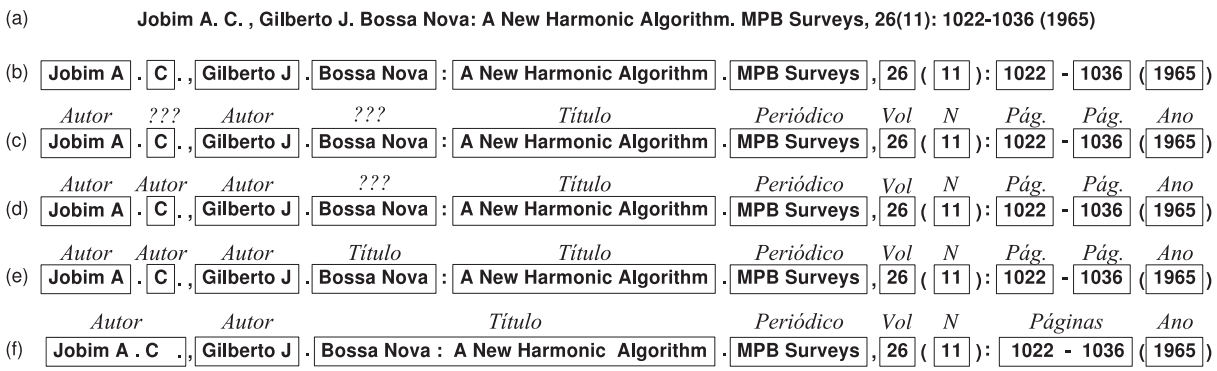


Figura 2.2: Exemplo de uma citação (a) e cada passo da extração: blocking (b), matching (c), binding (d, e), e joining (f).

### 2.1.3 p-delimitador

Um *p-delimitador*, ou *delimitador em potencial* é qualquer caractere distinto de **A**,...,**Z**, **a**, ...,**z**, **0**,...,**9**. Note que o método não assume que os p-delimitadores delimitam os campos bibliográficos. Ao invés disso, como explicado a seguir, analisamos cada um deles para verificar se realmente são usados como delimitadores de campo na citação que está sendo processada.

Note que a existência de delimitadores em potencial é essencial para a correta execução de nosso método, uma vez estes são utilizados estes para delimitar blocos que futuramente serão associados a um dado campo bibliográfico. Como exemplo de potenciais delimitadores podemos citar o conjunto $P = (“,”, “.”, “ : ”, “;”)$ que foi obtido a partir da coleção de Ciências da Saúde.

## 2.2 Fases do Método

A Figura 2.3 apresenta a visão geral do método de extração FLUX-CiM. A seguir, discutimos os detalhes de cada fase envolvida no processo de extração.

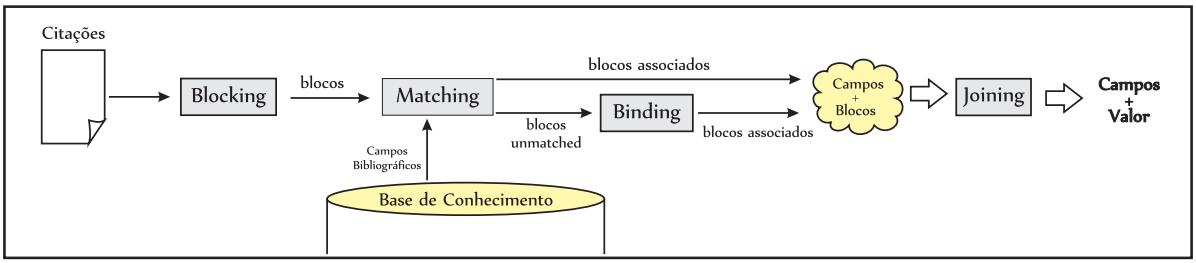


Figura 2.3: FLUX-CiM: Visão Geral

### 2.2.1 Blocking

A primeira fase do nosso método de extração, consiste em dividir uma citação em *substrings* que chamamos de *blocos*. Sejam $p_l$ e $p_r$ p-delimitadores e $C$ uma citação. Um bloco $b$ é uma *substring* de $C$ tal que:

- ocorre na seqüência $p_l b p_r$ e não contem nenhum p-delimitador, ou
- ocorre na seqüência $b p_r$ onde $b$ é o prefixo de $C$, ou
- ocorre na seqüência $p_l b$ onde $b$ é o sufixo de $C$.

Em nosso método, consideramos cada bloco como conjunto de termos que irão compor um valor de um determinado campo bibliográfico. Em uma mesma citação pode haver mais de um bloco que será associado a um mesmo campo. Na Figura 2.2(b) os blocos identificados para o nosso exemplo estão marcados com retângulos. O princípio subjacente à idéia de identificação dos blocos é a observação de que, geralmente, em uma citação cada valor de um campo bibliográfico é delimitado por um p-delimitador, mas nem todos os p-delimitadores delimitam um campo. Esse importante princípio foi avaliado experimentalmente, e os resultados são apresentados nessa dissertação(Seção 3.2.1).

### 2.2.2 Matching

A etapa de *Matching* consiste em associar cada bloco a um campo de metadados bibliográficos. Para realizar isto, comparamos cada bloco com as ocorrências que compõem a base de conhecimento e avaliamos a qual campo bibliográfico o bloco é mais provável

pertencer. Para alguns casos esta tarefa é simples de realizar. Por exemplo, o termo "procedure" é claramente relacionado ao campo *Título*. Em outros casos, existem termos ambíguos e precisamos utilizar as ocorrências para estimar o grau de ambiguidade de termos no que diz respeito aos campos bibliográficos da base de conhecimento.

Por exemplo, considere a base de conhecimento da Figura 2.1. De acordo com estas ocorrências, o termo **Potter** é considerado ambíguo, pois é encontrado em ocorrências dos campos *Autor* e *Título*. Por outro lado, o termo **Halloween** é uma típica ocorrência do campo *Título*, portanto, não ambíguo.

Na fase de matching, valores textuais (por exemplo título, nome de autor, etc.) são manipulados utilizando uma função de similaridade que chamamos *FF* (Field Frequency), que é uma adaptação da função AF, proposta em [Mesquita et al., 2007]. A função FF é definida a seguir:

$$FF(b, m_i) = \frac{\sum_{t \in T(m_i) \cap T(b)} fitness(t, m_i)}{|T(b)|} \tag{2.1}$$

onde $T(m_i)$ é o conjunto de todos os termos encontrados nas ocorrências do campo bibliográfico $m_i$, e $T(b)$ é o conjunto de termos presentes no bloco $b$. A função $fitness(t, m_i)$ computa a medida de *fitness* (descrita a seguir).

A função FF estima a probabilidade de $b$ fazer parte de uma ocorrência do campo bibliográfico $m_i$, através da avaliação de quão típicos os termos de $b$ são em relação as ocorrências deste campo de acordo com a base de conhecimento. Para isso, a medida de *fitness* é definida a seguir.

Dado um termo ambíguo, a função de **fitness** mede o quão típico um termo é em cada campo bibliográfico em que o mesmo ocorre. Por exemplo, nas ocorrências da base de conhecimento da Figura 2.1, o termo ambíguo **Potter** é mais típico no campo bibliográfico *Título* do que no campo *Autor*.

A função de **fitness** é computada de acordo com a seguinte fórmula:

$$fitness(t, m_i) = \frac{f(t, m_i)}{N(t)} \times \frac{f(t, m_i)}{f_{max}(m_i)} \tag{2.2}$$

onde $f(t, m_i)$ é o número de ocorrências $o_{i,k} \in O_i$ associadas com o campo bibliográfico $m_i$ que contem o termo $t$ na base de conhecimento, $f_{max}(m_i)$ é a maior freqüência de um termo entre todas as ocorrências $o_{i,k} \in O_i$, e $N(t)$ é o número total de ocorrências do termo $t$ na base de conhecimento.

A primeira fração da Equação 2.2 expressa a probabilidade do termo $t$ fazer parte de uma ocorrência de $m_i$. Essa probabilidade seria adequada para os nossos propósitos se todos campos $m_i$ tivessem o mesmo número de ocorrências na base de conhecimento. Porém, em geral isso não é verdade, pois campos com mais ocorrências tendem a ter valores maiores de probabilidade. Por isso, adicionamos a segunda fração, como um fator de normalização para evitar tal problema. Esta fração apresenta a freqüência de $t$ nas ocorrências de $m_i$ normalizado pela maior freqüência de um termo nas ocorrências de $m_i$. Assim, varia de 0, o que significa completamente raro, a 1, o que significa que esta é a mais freqüente. Essa normalização é também útil para fazer com que aos valores de freqüência sejam comparáveis entre todos os campos bibliográficos.

Desta forma, para cada bloco $b$ na citação, calculamos $FF(m_i, b)$, para cada campo $m_i$ na base de conhecimento. Por fim, $b$ é associado ao campo que alcança o valor máximo de $FF$.

Para o caso de valores numéricos (por exemplo, números de página, ano, volume, etc) tradicionais funções de similaridade textual não funcionam corretamente [Agrawal et al., 2003]. Assim, para atributos numéricos, utilizamos uma outra abordagem simples, mas eficaz: assumimos que os valores em cada campo bibliográfico seguem uma distribuição gaussiana. A similaridade entre o valor presente na citação e os valores da BC é definida como o valor médio da função densidade de probabilidade. Chamamos esta função *NM* (*Numeric Matching*). Esta função é normalizada utilizando-se a densidade da probabilidade máxima, que é alcançada quando um determinado valor é igual à média. Assim, definimos o valor de similaridade para valores numéricos da seguinte maneira:

$$NM(b, m_i) = \frac{1}{|b|} \sum_{v \in b} e^{-} \frac{v - \mu}{2\sigma^2} \tag{2.3}$$

onde $\sigma$ e $\mu$ são o desvio padrão e a média, respectivamente, dos valores do campo bibliográfico $m_i$.

Após a fase de matching, a maioria dos blocos estão associados a um dos campos bibliográficos da base de conhecimento. Referimo-nos a estes blocos como *matched*. No entanto, ainda podem ocorrer blocos *unmatched*, ou seja, alguns blocos podem permanecer sem associação com qualquer campo após a fase de matching. Esta situação ocorre com blocos compostos por termos não presentes entre as ocorrências da base de conhecimento.

Na Figura 2.2(c) exemplificamos a saída da fase de matching. Nesta figura, blocos *unmatched* são marcados com *???* e blocos *matched* são marcados com os nomes dos seus respectivos campos bibliográficos. Casos como esses devem ser solucionados, e tal tarefa é realizada pela fase de *binding*, que é explicada a seguir.

### 2.2.3 Binding

Na fase de matching, vários blocos foram associados a um campo bibliográfico de acordo com a base de conhecimento. Com base nesta informação, a fase de binding associa os blocos unmatched restantes com campos bibliográficos. Na Figura 2.2(c), ilustramos dois casos de blocos unmatched (marcadas com *???*). No entanto, em geral, poderia haver uma seqüência de blocos unmatched que precisam ser associados a algum campo. A nossa forma de resolver este problema depende da vizinhança da seqüência de blocos unmatched da citação. Há três casos distintos que consideramos: *vizinhança homogênea*, *vizinhança parcial* e *vizinhança heterogênea*. Para cada um destes casos, detalhamos abaixo a estratégia específica de binding que foi adotada.

#### Vizinhança Homogênea

Sejam $l$ e $r$ blocos associados com o mesmo campo bibliográfico $m$. Suponha que estes blocos ocorram na seqüência $l, p_0, u_1, p_1, \ldots, u_n, p_n, r$, onde cada $u_i$ é um bloco unmatched e cada $p_i$ é um p-delimitador. Neste caso, todos $u_i$ são associados ao campo $m$. Um exemplo de vizinhança homogênea é ilustrado na Figura 2.2(c), onde o bloco contendo o

termo "C"é associado a *Autor* na Figura 2.2(d) dado que sua vizinhança está associado a este campo bibliográfico.

### Vizinhança Parcial

Seja $b$ um bloco associado com o campo bibliográfico $m$. Suponha que este bloco ocorre na seqüência $I = u_1, p_1, \ldots, u_n, p_n, b$ ou na seqüência $F = b, p_0, u_1, p_1, \ldots, u_n$, onde cada $u_i$ é um bloco unmatched e cada $p_i$ é um p-delimitador. Neste caso, todos $u_i$ são associados com ao campo $m$. Note que em $I$, blocos $u_i$ iniciam a citação, enquanto em $F$, blocos $u_i$ terminam a mesma.

### Vizinhança Heterogênea

Considere o exemplo na Figura 2.2(c), onde temos que decidir se o bloco contendo "Bossa Nova" deve ser associado a *Autor*, como o bloco de esquerda, ou a *Título* como o bloco da direita.

Em tais situações, nosso método recorre aos p-delimitadores que ocupam a vizinhança do bloco unmatched. Verifica-se se estes p-delimitadores (1) são tipicamente encontrados entre blocos contíguos de campos distintos, ou (2) se são tipicamente encontradas entre blocos contíguos de um mesmo campo. No primeiro caso, consideramos que o p-delimitador é na verdade um delimitador de campo e, assim, os dois blocos por ele separados não podem estar associados ao mesmo campo bibliográfico. No segundo caso, consideramos que o p-delimitador é simplesmente um caractere que aparece nos valores de um campo e, assim, é provável que os dois blocos devam ser associados a um mesmo campo. Esta verificação é realizada com base nos resultados da fase de matching para um conjunto de citações, onde vários blocos estão rotulados com o seu campo correspondente. Assim, podemos analisar o quão comum um p-delimitador é para cada campo e como se comportam normalmente, ou seja, qual dos casos (1) ou (2), descritos acima, devem ser aplicados.

Por exemplo, na Figura 2.2, dado que "." é um provável delimitador entre os campos

*Autor* e *Título* e ":" é um provável caractere que ocorre nos valores do campo *Título*, associamos "Bossa Nova" ao campo *Título* ao invés de associar tal bloco ao campo *Autor*. Estas idéias são melhor elaboradas a seguir.

Considere a seqüência $l, p_0, u_1, p_1, \ldots, u_n, p_n, r$, onde $l$ e $r$ são blocos associados a campos bibliográficos distintos $m_l$ e $m_r$, respectivamente, $u_i$ são blocos unmatched e $p_i$ são p-delimitadores. Nosso problema é determinar, para cada $u_i$, se este deverá ser associado ao campo $m_l$ ou ao campo $m_r$. Primeiramente, consideramos que somente um dos p-delimitadores $p_i$ é verdadeiramente um delimitador entre campos bibliográficos.

Baseado nisso, uma vez que encontramos que algum $p_i$ é um delimitador de campo, então associamos todos os blocos unmatched $u_j$ $(0 < j \leq i)$ a $m_l$, isto é, o mesmo campo que o bloco da esquerda, e associamos todos $u_k$ $(i > k \geq n)$ a $m_r$, isto é , o mesmo campo que o bloco da direita.

Agora, considere as seguintes expressões:

$$T(p_k, m_l, m_r) = \frac{f(p_k, m_l, m_r)}{\sum_{p_j \in P} f(p_j, m_l, m_r)} \tag{2.4}$$

onde $f(p, m_l, m_r)$ é a freqüência do p-delimitador $p$ entre blocos contíguos associados aos campos $m_l$ e $m_r$ pela fase de matching, e $P$ é o conjunto de todos os p-delimitadores.

$$C(p_k, m) = \frac{f(p_k, m)}{\sum_{p_j \in P} f(p_j, m)} \tag{2.5}$$

onde $f(p, m)$ é a freqüência do p-delimitador $p$ entre blocos contíguos associados ao mesmo campo $m$ pela fase de matching, e $P$ é o conjunto de todos os p-delimitadores.

Intuitivamente, a Equação 2.4 estima a probabilidade de um dado p-delimitador $p_i$ ser um delimitador entre os campos bibliográficos $m_l$ e $m_r$, enquanto a Equação 2.5 estima a probabilidade de $p_i$ ser um caractere que ocorre como parte dos valores do campo $m$. É importante notar que as freqüências utilizadas nestas equações são obtidas após a análise de cada p-delimitador em todas as citações que devem ser extraídas. Isto é feito para assegurar que estatísticas representativas a cerca do posicionamento de cada

p-delimitador estão sendo produzidas.

Em nosso método, estes fatores são considerados para decidirmos qual p-delimitador $p_i$ é o delimitador entre os campos na citação. Para isso, utilizamos a Equação 2.6, definida a seguir.

$$\begin{aligned} D(p_k, m_l, m_r) = \quad & 1 - [(1 - T(p_k, m_l, m_r) \quad \times \\ & \prod_{0 \leq j < k} 1 - C(p_j, m_l) \quad \times \\ & \prod_{k > j \geq n} 1 - C(p_j, m_r)] \end{aligned} \tag{2.6}$$

onde $p_k$ é um p-delimitador e $k$ é a sua posição ordinal.

Dado um delimitador $p_k$, a Equação 2.6 leva em consideração: (1) a probabilidade de $p_k$ ser um delimitador típico de campo entre os valores de $m_l$ e $m_r$; (2) a probabilidade dos p-delimitadores a esquerda de $p_k$ fazerem parte dos valores do campo $m_l$; e (3) a probabilidade dos p-delimitadores a direita de $p_k$ fazerem parte dos valores do campo $m_r$.

Desta forma, o problema de associar uma seqüência de blocos unmatched que acontecem em uma vizinhança heterogênea é solucionado calculando-se $D(p_k, m_l, m_r)$ para cada p-delimitador $p_k$ presente na seqüência. O delimitador de campo é selecionado de acordo com maior valor atingido por esta equação.

Na Figura 2.2(e), por exemplo, o bloco contendo o termo "Bossa Nova" é então associado ao campo *Título*, dado que $D(\text{“ : ”}, Titulo, Autor) < D(\text{“.”}, Titulo, Autor)$.

### 2.2.4 Joining

Quando a fase de binding é finalizada, cada bloco na citação está associado a um campo de metadados. Em seguida, o último passo em nosso método de extração consiste em juntar blocos associados a um mesmo campo com o intuito de formar os valores do campo. Para a maioria dos casos, este passo é simples de realizar, uma vez que requer simplesmente juntar blocos contíguos associados a um mesmo campo bibliográfico. No entanto, juntar

blocos associados ao campo *Autor* requer um procedimento mais cuidadoso, uma vez que podem existir vários valores para o campo *Autor* em uma citação. Assim, nesta seção, descrevemos a forma como tratamos os blocos para formar valores para o campo *Autor*. Por exemplo, os blocos do campo *Autor* na Figura 2.2(e), devem ser unidos, para formar os valores de *Autor* como é ilustrado na Figura 2.2(f).

A solução para este problema utiliza informação disponível na base de conhecimento. Seja $\eta$ o número médio de termos nas ocorrências do campo *Autor* de acordo com a base de conhecimento. Assumimos que o número de termos encontrado nos valores de *Autor* em qualquer citação é aproximadamente igual a $\eta$.

Agora, considere um conjunto $s$ de caracteres utilizados como delimitadores implícitos para separar os valores do campo *Autor* em uma citação. Por exemplo, em uma dada citação, o caractere "," é utilizado como um delimitador para todos os valores de *Autor* a não ser para o último valor, o qual é separado pelo caractere "e". Neste caso, $s = \{$",","e"$\}$. No método FLUX-CiM, como mencionado anteriormente, observamos que o número de termos que é delimitado por estes caracteres em $s$ deve ser aproximadamente igual a $\eta$.

Considere a seqüência de blocos que deve ser unida para compor valores do campo *Autor*. Dado um conjunto de caracteres delimitadores $s$, dois ou mais blocos contíguos devem ser unidos se o p-delimitador $p$ entre eles não é um caractere delimitador, isto é, $p \notin s$. Portanto, devemos determinar quais são os p-delimitadores que compõem $s$.

A solução adotada é utilizar conjuntos delimitadores candidatos e para cada conjunto candidato, avaliarmos se este conjunto é o único que resulta em valores do campo *Autor* com o número de termos mais próximo a $\eta$. Para isso, definimos uma métrica que chamamos de *DE* (erro de delimitação) que é baseada na diferença entre os tamanhos dos valores (em número de termos) e o número médio de termos encontrados na base de conhecimento ($\eta$).

$$DE(s, a, \eta) = \prod_{x \in split(s,a)} dif(len(x), \eta) \tag{2.7}$$

onde $s$ é o conjunto de delimitadores, $a$ é a porção da citação composta por blocos

pertinentes ao campo *Autor*, e as seguintes funções auxiliares são utilizadas:

- $split(s, a)$ retorna todas as substrings de $a$ que são rodeadas por algum delimitador $p \in s$.
- $len(x)$ retorna o número de termos presentes em $x$.
- $dif(l_1, l_2) = |l_1 - l_2|$ se somente se $l_1 \neq l_2$, e $dif(l_1, l_2) = \epsilon_0$ se não, onde $\epsilon_0$ é uma constante mínima.

Intuitivamente, dada uma citação com um conjunto de blocos pertinentes ao campo *Autor* a serem unidos, a Equação 2.7 calcula um *score* baseado na distância entre $\eta$ e o número de termos de cada valor de *Autor* obtido quando se utiliza $s$ como o conjunto de delimitadores.

Assim, seja $P$ o conjunto de p-delimitadores entre os blocos do campo *Autor*. Avaliamos o erro de delimitação para cada sub-conjunto de p-delimitadores $s \subseteq P$ usando a Equação 2.7. O conjunto de delimitadores utilizados para os valores do campo *Autor* será o que obtém menor valor de erro de delimitação.

Como exemplo, considere a citação na Figura 2.2(e), na qual o conjunto de delimitadores entre os blocos do campo *Autor* é {“.”,“,”}. Além disso, assume-se $\eta = 2,7$ [2]. Quando o delimitador “,” é utilizado como um separador dos valores do campo *Autor*, o erro de delimitação é de cerca de 0,21, enquanto utilizando o delimitador “.” ou o conjunto de delimitadores {“.”,“,”}, o erro de delimitação é de cerca de 0,83 em ambos os casos. Assim, o delimitador “,” é a melhor escolha. Na Figura 2.2(f) mostramos valores do campo *Autor* obtidos com este delimitador.

## 2.3 Realimentação

Os experimentos realizados com o método de extração FLUX-CiM e que serão reportados no Capítulo 3, demonstram a elevada qualidade da extração e o elevado grau de flexibilidade do nosso método. No entanto, para que isto ocorra, é muito importante que a base

[2]Este é o valor real encontrado em uma das coleções de citação utilizada em nossos experimentos.

de conhecimento abranja uma parcela representativa do domínio de interesse. De acordo com nosso estudo, podemos afirmar que o tamanho da base de conhecimento influencia diretamente a qualidade da extração. Por outro lado, pode haver casos em que nova informação deve ser incorporada à base de conhecimento ao longo do tempo para refletir uma nova tendência encontrada no domínio. Por exemplo, o termo "Bluetooth" só foi recentemente incorporado ao vocabulário da área de informática. Esse fenômeno também pode ocorrer com valores de campos, como novos locais e nomes distintos de autores.

Para lidar com estes requisitos, ou seja, ter um número significativo de registros de citações na base de conhecimento e garantir a representatividade destas citações no que diz respeito ao estado atual de um determinado domínio, seria necessário coletar dados a partir desse domínio, por exemplo, da Internet, e adicionar esses dados à base de conhecimento, tal como descrito na Seção 2.1.1. Embora simples, para realizar tal procedimento é necessária a intervenção de um usuário, o que poderia tornar-se inconveniente em um cenário onde a autonomia é uma exigência.

Nesta seção propomos uma solução para este problema, através da incorporação direta dos resultados dos processos de extração na base de conhecimento, processo que chamamos de *Realimentação*, que é ilustrado na Figura 2.4. Considere uma base de conhecimento $K$ em um dado domínio $D$. Agora, suponha que usamos o método FLUX-CiM para extrair certa quantidade de citações a partir de um conjunto de teste $S$, também do domínio $D$. A realimentação consiste em tomar os termos que compõem os valores de dados extraídos de $S$ para cada campo e atualizar os campos correspondentes em $K$.

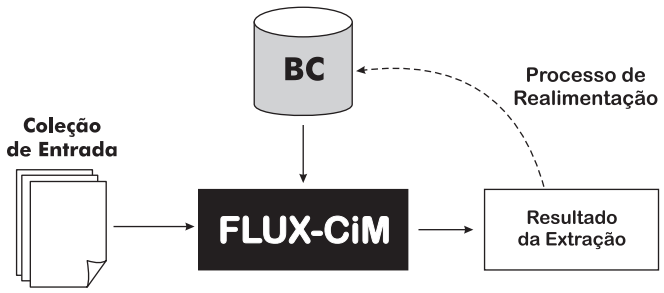


Figura 2.4: Processo de Realimentação

A primeira vista, utilizar diretamente os dados extraídos para atualizar a base de co-

nhecimento poderia introduzir uma certa quantidade de ruído (erros) a ela, o que poderia comprometer os resultados dos processos de extração baseado na base de conhecimento atualizada. No entanto, como é mostrado no Capítulo 3, o método FLUX-CiM realiza extração com alta qualidade, mesmo com uma pequena base de conhecimento. Assim, argumentamos que o uso dos resultados de extração para realizar o processo de realimentação é bastante "seguro", uma vez que a quantidade de ruído gerada é muito baixa.

Para corroborar esta afirmação, apresentamos resultados de experimentos que realizamos com o processo de realimentação reportados na Seção 3.4, do Capítulo 3.

# Capítulo 3

# Experimentos

Neste capítulo, apresentamos os experimentos realizados para avaliar a eficácia de nossa abordagem para a tarefa de extração de metadados de citações bibliográficas. Apresentamos também uma comparação experimental entre o nosso método e o CRF [Peng and McCallum, 2006], método considerado o estado da arte em extração de dados bibliográficos de artigos científicos. Por fim, apresentamos resultados de experimentos que realizamos com o processo de *Realimentação* para automaticamente expandir e atualizar a base de conhecimento de um dado domínio.

Em todos os experimentos foram realizadas tarefas de extração semelhantes sobre citações bibliográficas de três domínios distintos: *Ciências da Saúde* (CS1), *Ciências Sociais* (CS2) e *Ciência da Computação* (CORA). Em todos os casos, utilizamos amostras de registros de citações de cada domínio específico para gerar a base de conhecimento. Em seguida, executamos o método de extração sobre um conjunto de citações do mesmo domínio. A Tabela 3.1 apresenta algumas características de cada coleção que utilizamos em nossos experimentos. Observe que o número de campos de metadados da coleção CORA varia de 1 a 10. Isto acontece porque as citações desta coleção vêm de diferentes fontes, como artigos de conferências e revistas científicas de distintas editoras, e, portanto, elas têm estilos distintos.

| Domínio | Tamanho da BC | # Campos | # Citações |
|---|---|---|---|
| *CS1* | 5000 | 6 | 2000 |
| *CS2* | 5000 | 6 | 2000 |
| *CORA* | 350 | 1 a 10 | 150 |

Tabela 3.1: Características de cada coleção utilizada nos experimentos.

## 3.1 Configuração dos Experimentos

CORA é uma coleção heterogênea composta por 500 citações bibliográficas de várias conferências de ciência da computação e foi anteriormente utilizada em [Peng and McCallum, 2006] para avaliar o método CRF. Escolhemos aleatoriamente 350 citações para gerar a base de conhecimento para o nosso método e as outras 150 citações, foram utilizadas para teste. Esta proporção foi a mesma utilizada em [Peng and McCallum, 2006] para a avaliação do CRF.

Para os experimentos do domínio CS1, utilizamos uma coleção de citações da *PubMed Central (PMC)*[1]. No caso do domínio CS2, a coleção foi obtida a partir da Biblioteca Digital Scielo[2]. Para cada um desses domínios usamos uma coleções de mais de 50.000 registros de citações. As coleções CS1 e CS2 são ambas considerados *bem organizadas*, dado que as suas citações seguem um estilo uniforme, e *controladas*, uma vez que para cada citação, há uma seqüência estruturada de registros de metadados onde cada campo da citação é explicitamente identificado. Assim, através da realização de experimentos nestas coleções controladas podemos verificar, automaticamente, os resultados da extração para um grande volume de citações bibliográficas. As bases de conhecimento foram construídas com cinco mil registros de citações, enquanto o processo de extração utilizou outras duas mil citações distintas. Ao fazer isso, garantimos que não há sobreposição entre a base de conhecimento e o conjunto de teste.

As coleções CS1 e CS2 foram também utilizadas para a realização de um experimento que avalia como o nosso método se comporta quando o número de registros de citações na base de conhecimento varia. Para este experimento, variamos número de citações usados

[1] `http://www.pubmedcentral.nih.gov/`
[2] `http://www.scielo.org/`

para construir a BC de 50 até 10.000 registros de citações. Observe que não foi possível realizar este experimento com a coleção CORA devido ao pequeno número de citações disponíveis.

Todos os experimentos que relatamos neste capítulo foram repetidos cinco vezes. Assim, cada valor aqui apresentado representa a média dos valores obtidos em cada uma das cinco execuções.

Para a avaliação de nosso método, utilizamos as medidas: precisão, revocação e Medida F, na qual cada uma é computada a seguir. Seja $B_i$ o conjunto de referência e $S_i$ o conjunto de teste a ser comparado com $B_i$. Definimos precisão ($P_i$), revocação ($R_i$) e Medida F ($F_i$) como:

$$P_i = \frac{|B_i \cap S_i|}{|S_i|} \quad R_i = \frac{|B_i \cap S_i|}{|B_i|} \quad F_i = \frac{2(R_i.P_i)}{(R_i + P_i)} \tag{3.1}$$

## 3.2 Resultados

### 3.2.1 Hipótese de Blocking

O primeiro resultado que reportamos tem como objetivo verificar, na prática, a hipótese formulada quanto ao blocking, ou seja, que, geralmente em uma citação cada valor de um campo é delimitado por um p-delimitador, mas nem todos os p-delimitadores delimitam um valor. Para verificar isso, analisamos cada coleção de citações que é utilizada em nosso experimento e contamos os valores de campo que são delimitados por algum p-delimitador. Como esperado, em todos as coleções, 100% dos valores de um campo bibliográfico são delimitados por um p-delimitador.

### 3.2.2 Nível de blocos – Resultados

Apresentamos agora os resultados que mostram como os blocos foram corretamente associados com os seus respectivos campos através do nosso método.

Considere o conjunto de citações que utilizamos para avaliar o processo de extração em um determinado domínio. Seja $B_i$ o conjunto de todos os blocos nas citações neste conjunto que compõem os valores do campo de metadados bibliográficos $m_i$. Tais blocos foram utilizados como referências para a nossa verificação dos resultados a nível de blocos.

Agora, seja $S_i$ o conjunto de blocos associados a $m_i$ após uma determinado fase método FLUX-CiM, por exemplo, a fase de matching ou binding. A precisão e revocação obtidos com a Equação 3.1 para estes experimentos são apresentadas na Tabela 3.2 (a) para a coleção CORA, (b) para a coleção CS1 e (c) para CS2. Para comparar os resultados das duas primeiras etapas do nosso método, apresentamos separadamente os resultados obtidos após a etapa de matching e após o binding, os quais são cumulativos. Também apresentamos, o número de blocos unmatched após a fase de matching.

| Campo | Matching | | | Blocos | Binding | | |
|---|---|---|---|---|---|---|---|
| | P | R | F | Unmatched | P | R | F |
| *Autor* | 99,78% | 79,29% | 0,8836 | 20,63% | 99,82% | 98,96% | 0,9939 |
| *Título* | 98,11% | 90,43% | 0,9412 | 7,83% | 97,19% | 97,61% | 0,9740 |
| *Periódico* | 95,80% | 97,86% | 0,9682 | 1,43% | 95,80% | 97,86% | 0,9682 |
| *Data* | 99,70% | 97,38% | 0,9853 | 2,04% | 97,98% | 99,13% | 0,9855 |
| *Páginas* | 97,87% | 98,71% | 0,9829 | 1,29% | 97,06% | 99,14% | 0,9809 |
| *Conferência* | 100,00% | 96,00% | 0,9796 | 0,40% | 99,18% | 96,40% | 0,9777 |
| *Local* | 98,88% | 89,85% | 0,9415 | 9,64% | 98,48% | 98,48% | 0,9848 |
| *Editor* | 100,00% | 100,00% | 1,0000 | 0,00% | 100,00% | 100,00% | 1,0000 |
| *Número* | 97,87% | 97,87% | 0,9787 | 2,13% | 97,87% | 97,87% | 0,9787 |
| *Volume* | 100,00% | 98,25% | 0,9912 | 0,00% | 100,00% | 98,25% | 0,9912 |
| Média | 98,80% | 94,56% | 0,9652 | 4,54% | 98,34% | 98,37% | 0,9835 |

(a) CORA

| Campo | Matching | | | Unmatched | Binding | | |
|---|---|---|---|---|---|---|---|
| | P | R | F | Blocos | P | R | F |
| *Autor* | 99,04% | 94,33% | 0,9663 | 4,96% | 98,89% | 99,26% | 0,9907 |
| *Título* | 93,71% | 90,54% | 0,9210 | 6,17% | 92,90% | 95,96% | 0,9441 |
| *Periódico* | 97,51% | 89,22% | 0,9318 | 2,22% | 97,15% | 89,32% | 0,9307 |
| *Data* | 99,85% | 96,89% | 0,9835 | 0,00% | 99,85% | 96,89% | 0,9835 |
| *Páginas* | 99,90% | 98,54% | 0,9922 | 0,00% | 99,80% | 98,54% | 0,9917 |
| *Volume* | 98,53% | 97,65% | 0,9809 | 0,00% | 98,53% | 97,65% | 0,9809 |
| Média | 98,09% | 94,53% | 0,9626 | 2,22% | 97,86% | 96,27% | 0,9703 |

(b) CS1

| Campo | Matching | | | Unmatched | Binding | | |
|---|---|---|---|---|---|---|---|
| | P | R | F | Blocos | P | R | F |
| *Autor* | 99,35% | 95,26% | 0,9726 | 3,56% | 99,01% | 99,87% | 0,9044 |
| *Título* | 92,14% | 94,78% | 0,9344 | 5,89% | 91,17% | 98,43% | 0,9466 |
| *Periódico* | 98,22% | 94,41% | 0,9628 | 2,05% | 97,05% | 94,99% | 0,9601 |
| *Data* | 99,57% | 97,01% | 0,9827 | 0,00% | 99,57% | 99,01% | 0,9827 |
| *Páginas* | 99,65% | 98,45% | 0,9905 | 0,00% | 99,65% | 98,45% | 0,9905 |
| *Volume* | 98,67% | 98,66% | 0,9866 | 0,00% | 98,67% | 98,66% | 0,9866 |
| Média | 97,93% | 96,43% | 0,9716 | 1,91% | 97,52% | 97,90% | 0,9768 |

(c) CS2

Tabela 3.2: Precisão e revocação a nível dos blocos para cada campo após a fase de matching e binding para as coleções CORA (a), Ciências da Saúde (b) e Ciências Sociais (c), A porcentagem de blocos unmatched após a fase de matching também é apresentada.

No Capítulo 2 argumentamos que a fase de matching é a principal da nossa abordagem. Para averiguar isso, pode-se verificar que, em média, menos de 5% dos blocos são deixados unmatched para todos os conjuntos de citações. Isto ocorre porque blocos que apresentam pelo menos um dos seus termos com ocorrências na base de conhecimentos são associados a um campo bibliográfico. No entanto, este fator por si só não basta para garantir a alta precisão e revocação nos resultados obtidos, os quais são atribuidos à função FF (Equação 2.1) que propomos para realizar o matching.

Os resultados nas Tabelas 3.2 (a), (b) e (c) mostram ainda que a fase de binding desempenha um papel importante em nosso método, uma vez que foi capaz de melhorar significativamente os resultados de revocação mantendo níveis de precisão muito similares aos da fase de matching.

De forma geral, houve um único caso em que a fase de matching não foi capaz de distinguir os blocos entre dois campos bibliográficos distintos com elevada precisão. Isso ocorreu para os campos bibliográficos *Título* e *Periódico* da coleção CS1. Isto pode ser explicado pelo grande número de termos comuns entre estes dois campos.

Devemos salientar os elevados níveis de qualidade alcançados na coleção CORA, apesar do fato de esta coleção conter citações em diferentes estilos e o tamanho relativamente pequeno da base de conhecimento.

### 3.2.3 Nível de Campos – Resultados

Para demonstrar a eficácia do nosso método em todo processo de extração, avaliamos a qualidade da extração após a fase de joining, onde os blocos são unidos para compor os valores dos campos bibliográficos. Aqui, em vez de blocos, analisamos para cada campo que ocorre nas citações, se os valores atribuídos pelo nosso método a este campo estão corretos. Isto é muito importante, especialmente para o campo *Autor*, onde verificamos se os blocos associados a este campo foram corretamente unidos, ou seja, se os termos do mesmo nome de autor uniram-se para comportar o valor campo.

Neste caso, redefinimos a Equação 3.1, para considerar $B_i$ o conjunto completo de valores do campo $m_i$ e $S_i$ conjunto completo de valores associados a $m_i$ pelo nosso método. Novamente, os conjuntos $B_i$ foram automaticamente obtidos para todos as coleções. Os resultados são apresentados na Tabela 3.3 para as coleções CORA (a), CS1 (b) e CS2 (c). Observe que a precisão e a revocação são definidas aqui para valores de campo completos. Assim, se ao menos um bloco que compõe o valor do campo $m_i$ não for associado a $m_i$, consideramos que todos os valores de $m_i$ foram incorretamente extraídos.

Da Tabela 3.3 (a), (b) e (c) observamos que os elevados níveis precisão alcançados após a fase de matching e binding permanecem após a fase de joining. A exceção foi o valor de medida F para o campo *Título* do domínio de Ciências da Saúde, que foi de cerca de $0,85$. Analisando mais atentamente os valores do presente campo pode-se observar uma grande sobreposição com os termos de valores de campo *Periódico* neste domínio. Por este motivo, alguns blocos do campo *Periódico* foram erroneamente associados ao campo *Título* na fase de matching. Isto pode ser observado ao olharmos para o valores de revocação para o campo *Periódico* (89,32%) e de precisão para o campo *Título* (93,7%), após a fase de matching na Tabela 3.2(b), que são relativamente baixos. Esta situação foi propagada através da fase de binding até a fase de joining.

| Campo | Precisão | Revocação | Medida F |
|---|---|---|---|
| *Autor* | 97,21% | 98,67% | 0,9793 |
| *Título* | 93,01% | 96,67% | 0,9480 |
| *Periódico* | 93,45% | 91,5% | 0,9246 |
| *Data* | 96,01% | 90,01% | 0,9291 |
| *Páginas* | 97,98% | 98,81% | 0,9839 |
| *Conferência* | 93,16% | 91,73% | 0,9244 |
| *Local* | 90,01% | 94,78% | 0,9233 |
| *Editor* | 90,76% | 91,66% | 0,9121 |
| *Número* | 94,67% | 91,9% | 0,9326 |
| *Volume* | 100,00% | 99,14% | 0,9957 |
| Média | 96,28% | 95,8% | 0,9601 |

(a) CORA

| Campo | Precisão | Revocação | Medida F |
|---|---|---|---|
| *Autor* | 98,57% | 99,04% | 0,9880 |
| *Título* | 84,88% | 85,14% | 0,8501 |
| *Periódico* | 97,23% | 89,35% | 0,9312 |
| *Data* | 99,85% | 99,50% | 0,9967 |
| *Páginas* | 99,70% | 99,20% | 0,9945 |
| *Volume* | 96,41% | 98,75% | 0,9757 |
| Média | 96,11% | 95,16% | 0,9560 |

(b) CS1

| Campo | Precisão | Revocação | Medida F |
|---|---|---|---|
| *Autor* | 96,48% | 99,17% | 0,9781 |
| *Título* | 91,20% | 96,67% | 0,9386 |
| *Periódico* | 97,99% | 93,68% | 0,9579 |
| *Data* | 99,57% | 97,01% | 0,9827 |
| *Páginas* | 99,65% | 98,45% | 0,9905 |
| *Volume* | 98,67% | 98,66% | 0,9866 |
| Média | 97,26% | 97,27% | 0,9724 |

(c) CS2

Tabela 3.3: Valores de Precisão Revocação para cada campo após a fase de joining para as coleções CORA (a), CS1 (b) e CS2 (c).

### 3.2.4 Nível de Citações – Resultados

O último aspecto que analisamos em nossos experimentos é o quão bem cada registro de citação foi extraído pelo nosso método, ou seja, queremos verificar se os campos que compõem cada registro foram extraídos corretamente ou não. Observe que embora os resultados ao nível dos campos apresentados acima envolvem todos os valores a partir de um determinado campo, independentemente das citações em que ocorrem, nesta seção

analisamos a extração resultante em cada citação.

Para apresentar esses resultados, considere cada conjunto de referência $B_i$ como o conjunto de valores de campo em um dado registro de citação $C_i$. Agora, seja $S_i$ o conjunto de valores de campo extraídos de $C_i$ pelo nosso método. Então, a precisão e a revocação são calculadas através da equação 3.1. Na Tabela 3.4 apresentamos as médias de precisão e revocação obtidas nos experimentos para todos as coleções.

| Domínio | Precisão | Revocação | Medida F |
|---|---|---|---|
| *CS1* | 94,82% | 95,10% | 0,9496 |
| *CS2* | 97,32% | 97,21% | 0,9726 |
| *CORA* | 92,14% | 94,78% | 0,9344 |

Tabela 3.4: Valores de precisão e revocação para as citações após a fase de joining.

Os dados na Tabela 3.4 foram obtidos levando em consideração todos os valores de campos que ocorrem em cada citação, o que pode variar para cada cada citação individual. Estes resultados demonstram que o nosso método é capaz de lidar com uma variedade de tipos de citação, sem ter de depender de um conjunto pré-definido de estilos de citação.

Em nosso experimento final desta seção, verificamos como nosso método funciona quando o tamanho da base de conhecimento varia. O resultado deste experimento é apresentado na Figura 3.1, em que para os domínios CS1 e CS2, usamos um número crescente de amostra registros de metadados das citações, a partir de 50 a 10.000, e

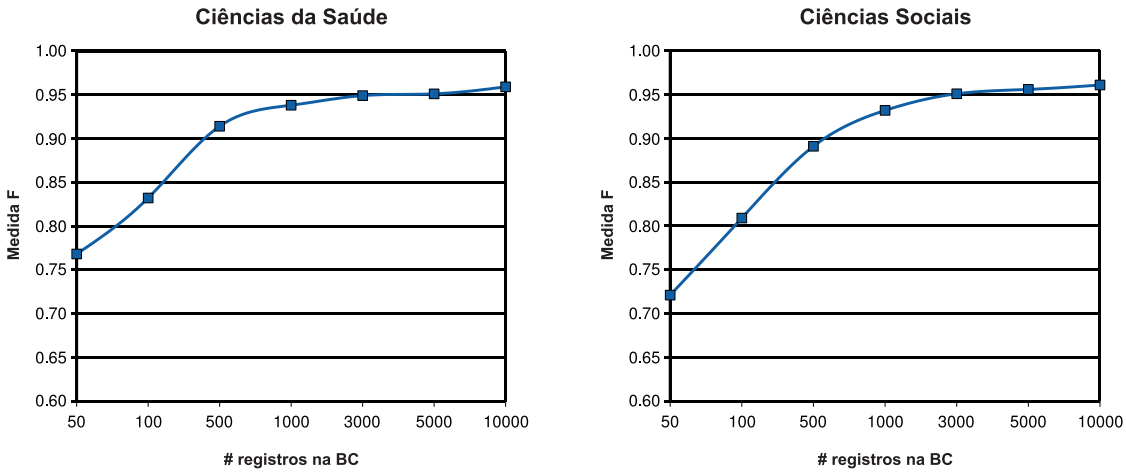


Figura 3.1: Desempenho da extração de citações relativa ao tamanho da base de conhecimento para os domínios de Ciências da Saúde e Ciências Sociais.

calculamos a medida F pra cada citação, resultantes da execução do processo de extração ao longo de cada coleção. Observe que em ambos os casos, os valores de medida F rapidamente se estabilizam, alcançando mais de 0,95 com 3.000 registros na base de conhecimento, e este valor continua a ser a mesmo quando a amostra de registros na base de conhecimento é de até 10.000. Isto mostra que nosso método não requer uma grande base de conhecimento para alcançar uma boa qualidade na extração nas coleções CS1 e CS2 que utilizamos. Como já mencionado, não foi possível realizar este experimento com a coleção CORA devido a o pequeno número de citações disponíveis nessa coleção.

### 3.2.5 Comentários Gerais

Embora os resultados experimentais demonstram a alta eficácia do nosso método, o problema da extração em citações é ainda um desafio, principalmente devido a alguns casos patológicos que impedem qualquer método de alcançar um resultado perfeito.

Na Figura 3.2, apresentamos 2 exemplos reais de citações bibliográficas e seus respectivos resultados produzidos pela extração com o FLUX-CiM. Na primeira citação, Figura 3.2(a), um dos valores do campo *Autor* é "Pathology Review Committee", que foi equivocadamente identificado no processo de extração com um valor do campo *Título*, uma vez que o termo "Pathology" é típico neste campo bibliográfico.

Na citação bibliográfica apresentado na Figura 3.2(b), é difícil até mesmo para os seres humanos separar os nomes dos autores corretamente. Procurando outras citações do mesmo artigo e descobriu-se que os valores corretos são: "Clayton Lewis", "Charles D. Hair" e "Victor Schoenberg". Note que o primeiro valor foi representado de forma distinta dos outros dois.

| | Citação Bibliográfica | Resultado da Extração |
|---|---|---|
| (a) | Nagtegaal ID, Klein Kranenbarg E, Hermans J, van de Velde CJH, van Krieken JHJM, Pathology Review Committee. Pathology data in the central database of multicenter randomized trials need to be based on pathology reports and controlled by trained quality managers. J Clin Oncol. 2000;18:1771-1779. | ***Autor 1***: Nagtegaal ID<br>***Autor 2***: Klein Kranenbarg E<br>***Autor 3***: Hermans J<br>***Autor 4***: van de Velde CJH<br>***Autor 5***: van Krieken JHJM<br>***Título***: Pathology Review Committee. Pathology data in the central database of multicenter randomized trials need to be based on pathology reports and controlled by trained quality managers.<br>***Periódico:*** J Clin Oncol.<br>***Ano***: 2000<br>***Volume***: 18<br>***Páginas***: 1771-1779. |
| (b) | Lewis, Clayton, D. Charles Hair, Victor Schoenberg (1989). Generalization Consistency Control. In Proceedings of ACM CHI'89 Conference on Human Factors in Computing Systems. pages 1-5. | ***Autor 1***: Lewis, Clayton, D<br>***Autor 2***: Charles Hair<br>***Autor 3***: Victor Schoenberg<br>***Ano***: 1989<br>***Título***: Generalization Consistency Control<br>***Conferência:*** In Proceedings of ACM CHI'89 Conference on Human Factors in Computing Systems<br>***Páginas***: pages 1-5. |

Figura 3.2: Exemplos de casos patológicos de citações bibliográficas da coleção CS1 (a) and CORA (b) and e seus respectivos resultados de extração.

## 3.3 Comparação entre FLUX-CiM e CRF

### 3.3.1 Comparação Experimental

Apresentamos agora os resultados da comparação experimental que realizamos entre FLUX-CiM e CRF, o método do estado da arte para extração de citações. Observamos que a saída fornecida pelo CRF é ligeiramente diferente da saída fornecida pelo FLUX-CiM, no sentido de que os valores em campos multivalorados, tais como nomes de autor, não são separados individualmente. Assim, para assegurar uma comparação justa entre os dois métodos, os resultados aqui expressos estão de acordo com as mesmas métricas utilizadas em [Peng and McCallum, 2006]. Para isso, redefinimos a Equação 3.1 para considerar $B_i$ como o conjunto de valores corretos de um campo $m_i$ e $S_i$ como o conjunto de valores associados a $m_i$ por um dado método de extração.

Para todas as três coleções, executamos a implementação de CRF publicamente disponível [3], que foi implementada de acordo com [Lafferty et al., 2001]. Para assegurar uma comparação justa, o mesmo conjunto de registros de citações foram usados para treinar o

[3] http://crf.sourceforge.net

| Campo | FLUX-CiM | CRF | Teste-T | Wilcoxon |
|---|---|---|---|---|
| *Autor* | 0,9420 | 0,9940 | - | - |
| *Título* | 0,9357 | **0,9830** | 2,00% | 2,00% |
| *Periódico* | **0,9262** | 0,9130 | 1,00% | 1,00% |
| *Data* | 0,9566 | **0,9890** | 3,00% | 5,00% |
| *Páginas* | 0,9567 | 0,9860 | - | - |
| *Conferência* | 0,9364 | 0,9370 | - | - |
| *Local* | **0,9315** | 0,8720 | 1,00% | 1,00% |
| *Editor* | **0,9250** | 0,7610 | 1,00% | 1,00% |
| *Número* | **0,9408** | 0,8940 | 1,00% | 1,00% |
| *Volume* | **0,9995** | 0,9592 | 1,00% | 1,00% |
| Média | **0,9390** | 0,9254 | 3,00% | 1,00% |

(a) CORA

| Campo | FLUX-CiM | CRF | Teste-T | Wilcoxon |
|---|---|---|---|---|
| *Autor* | **0,9662** | 0,9548 | 4,00% | 2,00% |
| *Título* | **0,9956** | 0,9616 | 1,00% | 1,00% |
| *Periódico* | **0,9371** | 0,8930 | 1,00% | 1,00% |
| *Data* | **0,9987** | 0,9657 | 2,00% | 2,00% |
| *Páginas* | 0,9783 | 0,9647 | - | - |
| *Volume* | **0,9995** | 0,9592 | 1,00% | 1,00% |
| Média | **0,9792** | 0,9498 | 1,00% | 1,00% |

(b) CS1

| Campo | FLUX-CiM | CRF | Teste-T | Wilcoxon |
|---|---|---|---|---|
| *Autor* | **0,9954** | 0,9431 | 1,00% | 1,00% |
| *Título* | **0,9978** | 0,9714 | 1,00% | 1,00% |
| *Periódico* | **0,9401** | 0,8889 | 1,00% | 1,00% |
| *Data* | **0,9984** | 0,9619 | 3,00% | 5,00% |
| *Páginas* | **0,9318** | 0,9067 | 1,00% | 1,00% |
| *Volume* | **0,9720** | 0,9214 | 1,00% | 1,00% |
| Média | **0,9726** | 0,9322 | 1,00% | 1,00% |

(c) CS2

Tabela 3.5: Resultados comparativos de medida F para as coleções CORA (a), CS1 (b) e CS2 (c).

modelo para o CRF e para gerar a base de conhecimento para o FLUX-CiM. Do mesmo modo, as mesmas citações no conjunto de teste foram aplicadas para ambos os métodos.

No caso das coleções CS1 e CS2, cada resultado apresentado foi obtido após 5 execuções completas , ou seja, em cada execução, um conjunto de treino para o CRF, uma BC para FLUX-CiM, e um conjunto de teste para ambos os métodos foram geradas aleatoriamente. Para o caso da coleção CORA, como em [Peng and McCallum, 2006], experimentos foram executados apenas uma vez, dado que o número de citações disponíveis nesta coleção é muito pequeno para permitir a não sobreposição nas execuções.

Para CS1 e CS2, utilizamos 5.000 citações para a formação de base de conhecimento e para o treino do CRF, e 2.000 citações para o teste. Para CORA, usamos uma base de conhecimento com 350 registros de citações, o mesmo número para o treino do CRF e, em seguida, 150 citações para testar ambos os métodos.

Para todas as comparações relatadas, foram utilizadas o teste Wilcoxon [Wilcoxon, 1945] e o Teste-T [Anderson and Finn, 1996] para determinar se a diferença no desempenho foi estatisticamente significativa. Em todos os casos, só tiramos conclusões a partir dos resultados obtidos que foram significativos em pelo menos 5% para ambos os testes.

Observando os resultados apresentados na Tabela 3.5, pode-se notar que em todas as três coleções o método FLUX-CiM alcança melhores resultados que o método CRF para a maioria dos campos, de acordo com ambos os testes estatísticos. Os melhores resultados obtidos pelo CRF nos dois campos da coleção CORA (em negrito) podem ser atribuídos ao número limitado de registros bibliográficos na base de conhecimento para essa coleção. Para as coleções CS1 e CS2, onde tínhamos um grande volume de dados bibliográficos para compormos os conjuntos de teste e a base de conhecimento, FLUX-CiM atuou melhor que o CRF para todos os campos. Estes experimentos demonstram que, mesmo sem nenhuma intervenção humana na criação de um conjunto de treino, FLUX-CiM alcança melhor qualidade na extração que o CRF, que é um método que necessita treinamento.

| | Exemplo de Estilo de Citação |
|---|---|
| *1* | Kerlikowske K, Orel SG, Troupin RH. Nonmammographic imaging. Semin Roentgenol. 1993;28:231-241 |
| *2* | 231-241: Nonmammographic imaging. Kerlikowske K: Orel SG: Troupin RH, 1993; 28. Semin Roentgenol |
| *3* | 1993; Kerlikowske K; Orel SG; Troupin RH; Semin Roentgenol. Nonmammographic imaging. 231-241: 28 |
| *4* | Nonmammographic imaging: 1993, Kerlikowske K, 231-241, Orel SG; Troupin RH. Semin Roentgeno |

(a)

| Conjunto | # de citações por estilo | # de citações no conjunto de teste |
|---|---|---|
| *1 estilo* | 2,000 | 2,000 |
| *2 estilos* | 1,000 | 2,000 |
| *3 estilos* | 667 | 2,001 |
| *4 estilos* | 500 | 2,000 |

(b)

Tabela 3.6: (a) Exemplos de estilos de citações e (b) Configuração final de cada conjunto de teste.

### 3.3.2 Lidando com diferentes estilos de apresentação nas citações

Como já discutido, uma das principais características que consideramos como muito importante no FLUX-CiM é a sua flexibilidade na extração de citações independentemente de um estilo utilizado. Isto acontece porque a nossa abordagem para extração não se baseia em padrões de codificação de delimitadores específicos utilizados em um estilo em particular, mas sim em características gerais das citações e nos valores de seus campos bibliográficos.

Para avaliar essa propriedade, realizamos uma série de experimentos em que os conjuntos de teste incluem citações com estilos distintos. Estes experimentos simulam situações em que citações, devem ser extraídas de vários artigos científicos a partir de diferentes fontes, com diferentes estilos de citações.

Nos experimentos, conjuntos de teste foram gerados da seguinte forma. Utilizamos citações das coleções CS1 e CS2 e geramos quatro conjuntos de teste, tal que o conjunto $i$ contém $\lceil N/i \rceil$ citações formatadas de acordo com o estilo $i$, onde $N = 2.000$ e $1 \geq i \leq 4$. Estilo 1 corresponde ao estilo original da citação usado em cada coleção. Os outros estilos foram aleatoriamente gerados através da mudança dos delimitadores de campo e da ordem relativa dos campos. Gerando estilos de citações de forma aleatória, visamos

simular situações onde citações com estilos não conhecidos anteriormente são utilizados.

Na Tabela 3.6 apresentamos (a) exemplos de estilos utilizados e (b) um resumo da configuração final de cada conjunto de teste. Para construir uma base de conhecimento para o FLUX-CiM e treinar o modelo do CRF escolhemos aleatoriamente 5.000 registros de citações em seu estilo original citações, ou seja, Estilo 1, de cada coleção respectiva

Os resultados deste experimento são apresentados na Tabela 3.7. Observe que, a medida F obtida com o CRF decresce com o aumento do número de estilos de citações. Isto acontece porque o modelo CRF baseia-se em características específicas aprendidas a partir de um único estilo em que foi treinado. Por outro lado, com o método FLUX-CiM a medida F permanece constante, independentemente do número de estilos de citações utilizado, corroborando, assim, as nossas hipóteses acerca da flexibilidade do nosso método.

| # de Estilos | FLUX-CiM | CRF | Wilcoxon | Teste-T |
|---|---|---|---|---|
| *1* | 0,9792 | 0,9498 | 1,00% | 1,00% |
| *2* | 0,9792 | 0,7065 | 1,00% | 1,00% |
| *3* | 0,9792 | 0,4033 | 1,00% | 1,00% |
| *4* | 0,9792 | 0,3567 | 1,00% | 1,00% |

(a) CS1

| # de Estilos | FLUX-CiM | CRF | Wilcoxon | Teste-T |
|---|---|---|---|---|
| *1* | 0,9704 | 0,9322 | 1,00% | 1,00% |
| *2* | 0,9704 | 0,7586 | 1,00% | 1,00% |
| *3* | 0,9704 | 0,3867 | 1,00% | 1,00% |
| *4* | 0,9704 | 0,3199 | 1,00% | 1,00% |

(b) CS2

Tabela 3.7: Valores de medida F obtidos com o uso de diferentes estilo de citações para as coleções CS1 (a) e CS2 (b).

## 3.4 Resultados do processo de Realimentação

Nos experimentos relatados a seguir, visamos mostrar a eficácia do processo de realimentação para automaticamente atualizar e expandir a base de conhecimento, sem comprometer a qualidade da extração. Conduzimos experimentos similares nos domínios de Ciências da Saúde e Ciências Sociais, uma vez que estes são os domínios para os quais

temos um grande conjunto de citações disponíveis. Nós analisamos o comportamento do nosso método usando a realimentação em 3 diferentes cenários. Primeiro, começamos com uma base de conhecimentos construída com apenas 50 registros de citações, e então executamos a tarefa de extração em um conjunto contendo 1.000 citações. Para cada execução, avaliamos a qualidade da extração em termos de medida F. Também realizamos experimentos começando com bases de conhecimento construídas com 1.000 e 3.000 registros de citações. Em todos os casos, os registros das citação para a base inicial e para a realimentação foram aleatoriamente escolhidos. Além disso, executamos cada experimento cinco vezes. Os resultados dos experimentos são apresentados na Figura 3.3, onde cada ponto representa a média dos valores obtidos em cada uma das cinco execuções.

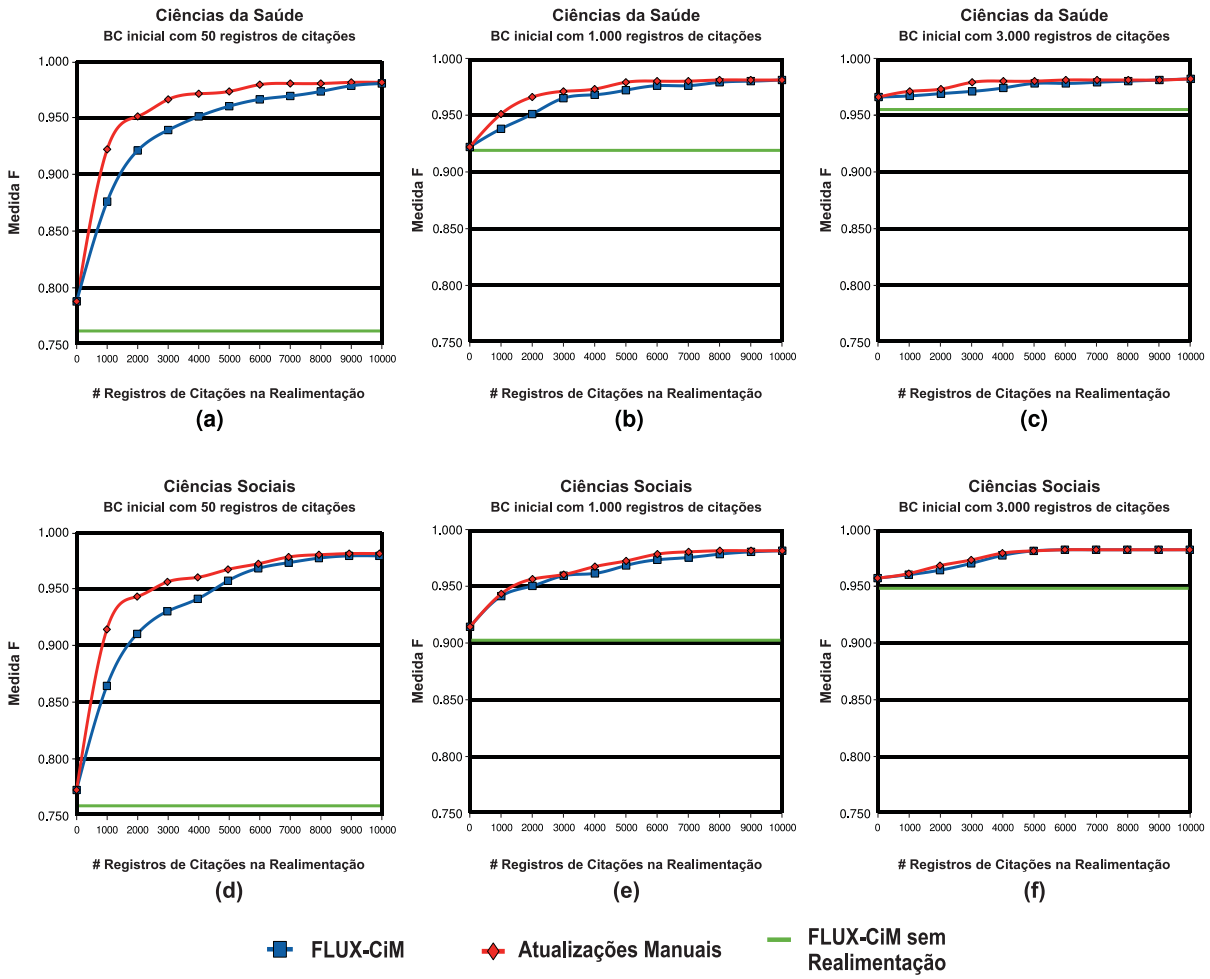


Figura 3.3: Desempenho da extração com realimentação nos domínios de Ciências da Saúde e Ciências Sociais.

Na Figura 3.3, cada gráfico mostra a qualidade alcançada pelo método FLUX-CiM em termos de medida F como uma função do número de registros de citações utilizados para atualizar a base de conhecimento, de acordo com o processo automático de realimentação.

Esta qualidade é comparada com o nível de qualidade que seria alcançado se a base de conhecimento fosse atualizada manualmente com o mesmo número de registros de citações totalmente corretos.

Para isso, usamos os registros de citações corretos correspondentes e adicionados à base de conhecimento. Isto representa os *Limites superiores* de qualidade que pode ser alcançado depois de usar a realimentação. A linha reta nestes gráficos representa a qualidade de extração alcançada se o processo fosse executado em todo o conjunto de teste, significa que, se o FLUX-CiM tivesse sido utilizado para realizar a extração sobre 10.000 citações com apenas a base de conhecimento inicial.

Como podemos notar, em todos os cenários distintos o processo de realimentação automática, com 9.000 registros de citações ou mais, atinge a qualidade dos níveis dos limites superiores. Isto significa que, mesmo quando se inicia com um pequeno conjunto de registros de citações na base de conhecimento, é possível utilizar o processo de realimentação de forma automática sem intervenção do usuário para atingir resultados de alta qualidade. Além disso, mesmo quando se inicia com uma pequena base de conhecimento, é melhor realizar a tarefa de extração em pequenos conjuntos de teste, do que em todo conjunto de citações disponível.

Nos gráficos da Figura 3.3 (c), (d), (e) e (f), em que a bases de conhecimento iniciais foram construídas utilizando 1.000 ou 3.000 registros de citações, o processo automático de realimentação traz a mesma melhoria na qualidade que seria obtida com as perfeitas atualizações manuais da base de conhecimento.

Estes resultados corroboram a nossa alegação de que, mesmo se nosso processo de realimentação vir a introduzir alguns erros na base de conhecimento, dado que a qualidade da extração obtida pelo método FLUX-CiM não é perfeita, é bom o suficiente para não comprometer operações de extração realizadas após a realimentação. Isto indica que a

atualização da base de conhecimento pode ser realizada automaticamente sem intervenção do usuário.

# Capítulo 4

# Conclusões e Trabalhos Futuros

Nesta dissertação, apresentamos um novo método, FLUX-CiM, para extrair componentes (por exemplo: nomes de autor, títulos de artigos, locais, números de página) de citações bibliográficas. Ao contrário dos métodos anteriores encontrados na literatura o nosso método não depende de padrões específicos para codificação de delimitadores utilizados em um determinado estilo citação. Esta característica nos proporciona um alto grau de automação e flexibilidade e permite que o método FLUX-CiM possa extrair componentes de referências bibliográficas em qualquer estilo de citação, como demonstrado pelos experimentos aqui reportados. FLUX-CIM utiliza uma base de conhecimento automaticamente construída a partir de um conjunto existente de registros de metadados de um determinado domínio (por exemplo: Ciência da Computação, Ciências da Saúde, Ciências Sociais, etc.). Em geral, estes registros podem ser facilmente encontrados na Web ou em outros repositórios públicos de dados.

O método FLUX-CiM difere de abordagens relacionadas, que dependem da construção manual das bases de conhecimento para então, reconhecer os componentes de uma citação. Além disso, FLUX-CIM funciona de maneira diferente dos métodos anteriores que se baseiam em modelos gerados através de treino guiados por um usuário. O processo de extração em nosso método baseia-se em: (1) estimar a probabilidade de um determinado termo encontrado em uma citação ocorrer como um valor de um determinado campo bibliográfico de acordo com as informações encontradas na base de conhecimento, e (2) a

utilização de propriedades estruturais genéricas presentes em citações bibliográficas.

A eficácia e a aplicabilidade do nosso método foram demonstradas por experimentos para extrair informações a partir de referências bibliográficas em artigos científicos de três domínios distintos: Ciências da Saúde (CS1), Ciência da Computação (CORA) e Ciências Sociais (CS2). Os experimentos realizados mostraram que FLUX-CiM obtém níveis de precisão e revocação superiores a 95% para os campos presentes no conjunto de citações e, revocação média de mais de 94% para os campos presentes em cada citação.

Realizamos também uma comparação experimental entre o método proposto e o método CRF. Os resultados desses experimentos demonstraram que, mesmo sem qualquer intervenção do usuário para criar um conjunto de treino, o método FLUX-CiM alcança melhor qualidade na extração do que o CRF.

A flexibilidade do FLUX-CiM foi experimentalmente verificada por meio de um conjunto de experimentos em que os conjuntos de teste incluíam citações com diferentes estilos. Com os resultados destes experimentos, corroboramos a nossa hipótese de que qualidade da extração permanece estável, independentemente do número de estilos de citações utilizados.

Por fim, propusemos um processo de *Realimentação* para automaticamente atualizar e expandir a base de conhecimento através da incorporação direta dos resultados de um processo de extração realizado pelo FLUX-CiM. Demonstramos através de experimentos que tal estratégia pode ser usada para alcançar bons resultados de extração, mesmo que apenas uma pequena amostra inicial de registros bibliográficos esteja disponível para a construção da base de conhecimento. Apesar da introdução de alguns erros na base de conhecimento por este processo, a qualidade dos resultados obtidos demonstra que este fato não compromete as operações de extração realizadas após a realimentação. Com isso, mostramos que, a atualização automática da base de conhecimento pode ser realizada sem intervenção de usuário através do FLUX-CiM.

O método aqui descrito foi publicado em três artigos [Cortez et al., 2007] [Cortez et al., 2009, Cortez and da Silva, 2008], os quais derivam deste trabalho de mes-

trado.

Como um trabalho futuro, pretende-se investigar diferentes funções de casamento aproximado que possam distinguir melhor os campos das citações que têm valores comuns para descrever seus domínios, por exemplo, nome do autor e nome do editor. Este tipo de função poderia tornar o nosso método mais geral e robusto.

Uma estratégia interessante para alcançar melhores resultados de extração em tipos de dados complexos seria a utilização do nosso método para descobrir automaticamente o estilo implícito dos dados e, então, usar essa propriedade como uma evidência em outro processo supervisionado de extração, por exemplo.

Consideramos também investigar a aplicabilidade do nosso método de extração de citações em outras fontes de citações além de artigos científicos. Por exemplo, parece ser interessante ter um mecanismo para preencher automaticamente uma Biblioteca Digital com metadados diretamente a partir de sites de conferências ou a partir dos cabeçalhos dos trabalhos publicados nestes locais.

# Referências Bibliográficas

[Agrawal et al., 2003] Agrawal, S., Chaudhuri, S., Das, G., and Gionis, A. (2003). Automated ranking of database query results. *Proceedings of CIDR 2003, Biennial Conference on Innovative Data Systems Research*.

[Anderson and Finn, 1996] Anderson, T. and Finn, J. (1996). *The New Statistical Analysis of Data*. Springer.

[Arasu and Garcia-Molina, 2003] Arasu, A. and Garcia-Molina, H. (2003). Extracting structured data from web pages. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 337–348, New York, NY, USA. ACM Press.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.

[Calado et al., 2006] Calado, P., Cristo, M., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., and Ziviani, N. (2006). Link-based similarity measures for the classification of web documents. *J. Am. Soc. Inf. Sci. Technol.*, 57(2):208–221.

[Cortez et al., 2007] Cortez, E., da Silva, A., Gonçalves, M., Mesquita, F., and de Moura, E. (2007). FLUX-CIM: flexible unsupervised extraction of citation metadata. *Proceedings of the 2007 Conference on Digital Libraries*, pages 215–224.

[Cortez et al., 2009] Cortez, E., da Silva, A., Gonçalves, M., Mesquita, F., and de Moura, E. (Online version, 2009). A flexible approach for extracting metadata from biblio-

graphic citations. *Journal of the American Society for Information Science and Technology.*

[Cortez and da Silva, 2008] Cortez, E. and da Silva, A. S. (2008). A flexible approach for extracting metadata from bibliographic citations. ACM SIGMOD/PODS Conference 2008. SIGMOD Undergraduate Posters.

[Couto et al., 2006] Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Ziviani, N., Moura, E., and Ribeiro-Neto, B. (2006). A comparative study of citations and links in document classification. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 75–84, New York, NY, USA. ACM Press.

[Crescenzi et al., 2001] Crescenzi, V., Mecca, G., and Merialdo, P. (2001). Roadrunner: Towards automatic data extraction from large web sites. In *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*, pages 109–118, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Culotta et al., 2006] Culotta, A., Kristjansson, T. T., McCallum, A., and Viola, P. A. (2006). Corrective feedback and persistent learning for information extraction. *Artif. Intell.*, 170(14-15):1101–1122.

[Day et al., 2005] Day, M.-Y., Tsai, T.-H., Sung, C.-L., Lee, C.-W., Wu, S.-H., Ong, C.-S., and Hsu, W.-L. (2005). A knowledge-based approach to citation extraction. In *IRI '05: Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration*, pages 50–55, New York, NY, USA. IEEE Systems, Man, and Cybernetics Society.

[Embley et al., 1999] Embley, D. W., Campbell, D. M., Jiang, Y. S., Liddle, S. W., Lonsdale, D. W., Ng, Y.-K., and Smith, R. D. (1999). Conceptual-model-based data extraction from multiple-record web pages. *Data Knowl. Eng.*, 31(3):227–251.

[Freitag and McCallum, 2000] Freitag, D. and McCallum, A. (2000). Information extraction with hmm structures learned by stochastic optimization. In *Proceedings of the*

*Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 584–589. AAAI Press / The MIT Press.

[Gonçalves et al., 2007] Gonçalves, M., Moreira, B., Fox, E., and Watson, L. (2007). "What is a good digital library?"–A quality model for digital libraries. *Information Processing and Management*, 43(5):1416–1437.

[Han et al., 2003] Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2003*, pages 37–48. IEEE Computer Society.

[Hsu and Dung, 1998] Hsu, C.-N. and Dung, M.-T. (1998). Generating finite-state transducers for semi-structured data extraction from the web. *Inf. Syst.*, 23(9):521–538.

[Hu et al., 2005] Hu, Y., Li, H., Cao, Y., Meyerzon, D., and Zheng, Q. (2005). Automatic extraction of titles from general documents using machine learning. In *JCDL'05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tools & techniques: supporting classification, pages 145–154.

[Kushmerick, 2000] Kushmerick, N. (2000). Wrapper induction: efficiency and expressiveness. *Artif. Intell.*, 118(1-2):15–68.

[Laender et al., 2002a] Laender, A. H. F., Ribeiro-Neto, B. A., and da Silva, A. S. (2002a). Debye - data extraction by example. *Data Knowl. Eng.*, 40(2):121–154.

[Laender et al., 2002b] Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. (2002b). A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93.

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

[Lawrence et al., 1999] Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, 32(6):67–71.

[Lee et al., 2007] Lee, D., Kang, J., Mitra, P., Giles, C. L., and On, B.-W. (2007). Are your citations clean? new scenarios and challenges in maintaining digital libraries. *Communications of the ACM*, 50(12):33–38.

[Liu et al., 2003] Liu, B., Grossman, R., and Zhai, Y. (2003). Mining data records in web pages. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–606, New York, NY, USA. ACM Press.

[Mesquita et al., 2007] Mesquita, F., da Silva, A., de Moura, E., Calado, P., and Laender, A. (2007). LABRADOR: Efficiently publishing relational databases on the web by using keyword-based query interfaces. *Information Processing and Management*, 43(4):983–1004.

[Muslea et al., 2001] Muslea, I., Minton, S., and Knoblock, C. A. (2001). Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1-2):93–114.

[OAI., 2005] OAI. (2005). The Open Archives Initiative protocol for metadata harvesting. *at http://www.openarchives.org/, accessed*, 26/10/2005.

[Paynter, 2005] Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CA, USA, June 7-11, 2005, Proceedings*, pages 291–300. ACM.

[Peng and McCallum, 2006] Peng, F. and McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Inf. Process. Manage.*, 42(4):963–979.

[Reis et al., 2004] Reis, D. C., Golgher, P. B., Silva, A. S., and Laender, A. F. (2004). Automatic web news extraction using tree edit distance. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 502–511, New York, NY, USA. ACM Press.

[Soderland, 1999] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83.

[Yilmazel et al., 2004] Yilmazel, O., Finneran, M., C., Liddy, and D., E. (2004). Metaextract: an NLP system to automatically assign metadata. In *JCDL'04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Collaboration and group work, pages 241–242.