Uma Abordagem Flexível para Extração de Metadados em Citações Bibliográficas

Eli Cortez[1], **Altigran Soares da Silva (Orientador)**[1]

[1]Departamento de Ciência da Computação – Universidade Federal do Amazonas (UFAM)
Manaus-AM, Brasil (Dept. onde dissertação foi aprovada)
{eccv,alti}@dcc.ufam.edu.br

Resumo. *Neste artigo apresentamos o FLUX-CiM, um novo método de extração de componentes de citações bibliográficas, tais como nomes de autores, títulos de artigo, etc. Tal método não se baseia em padrões específicos de codificação de delimitadores de um determinado estilo de citação, o que lhe confere um alto grau de automação e flexibilidade. Diferentemente de abordagens anteriores que dependem de treinamento manual para realizar o processo de extração, o nosso método necessita apenas de uma base de conhecimento que pode ser automaticamente construída a partir de um conjunto existente de registros de metadados de um dado domínio, por exemplo: Ciência da Computação, Ciências da Saúde, etc. Para demonstrar a eficácia e aplicabilidade do método proposto, realizamos experimentos que de extração dados de citações bibliográficas de artigos científicos. Os resultados destes experimentos 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, o FLUX-CiM produziu resultados superiores, sem a fase de treino que é exigida por esse método.*

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. Citações também têm sido utilizadas como fonte de evidências auxiliar em tarefas de Recuperação de Informação, tais como: classificação automática de documentos, classificação e avaliação da qualidade. 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 como: limpeza nos dados para correção de erros, verificação de atribuição imprópria de autoria, remoção de registros duplicados, etc.. 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. Porém, esta não é uma tarefa simples por inúmeras razões, tais como: erros na entrada de dados, variedade nos formatos de citação, nomes de autores ambíguos, além do grande volume e variedade de dados bibliográficos a considerar.

[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`

Nesta dissertação de mestrado abordamos o problema de extração de metadados em citações bibliográficas. Para isso, propomos uma novo método chamado *FLUX-CiM (**Fl**exible **U**nsupervised **Ex**traction – **Ci**tation **M**etada)* para ajudar a extrair corretamente componentes de citações bibliográficas. Diferentemente dos métodos tradicionais [Peng and McCallum 2006], que dependem de treinamento manual para realizar o reconhecimento de componentes em uma citação, nosso método necessita somente de uma base de conhecimento, que pode ser automaticamente construída a partir de um conjunto existente de registros de metadados de um dado domínio. Experimentos realizados para comparar nosso método, FLUX-CiM, com o CRF [Peng and McCallum 2006], o estado da arte em extração de informação, 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 aos obtidos pelo CRF quando o conjunto de teste possui citações formatadas com vários estilos diferentes.

Parte dos resultados da dissertação foram publicados nos anais da conferência ACM/IEE JCDL 2007 [Cortez et al. 2007] e no periódico internacional Journal of the American Society for Information Science and Technology (JASIST) [Cortez et al. 2009]. Além disso, uma ferramenta construída a partir do nosso método foi apresentada na Seção de Demos do SBBD 2008 [Cortez et al. 2008].

Este artigo está organizado da seguinte forma. Na Seção 2 apresentamos os principais trabalhos relacionados. Em seguida, na Seção 3 apresentamos o nosso método, introduzindo antes alguns conceitos básicos necessários para o seu entendimento. A Seção 4 descreve os experimentos realizados para verificar a qualidade do nosso método de extração em comparação com o estado da arte. Finalmente, na Seção 6 apresentamos as nossas conclusões e sugestões de trabalhos futuros. Por limitação de espaço, apenas alguns resultados principais foram transcritos neste documento. A descrição detalhada de nosso método assim como o detalhamento dos experimentos são apresentados no texto original da dissertação, disponível em `http://gtiexperimentos.com.br/elicortez/dissertacao-mestrado-elicortez.pdf`.

2. Trabalhos Relacionados

Nos últimos anos, várias ferramentas, métodos e técnicas têm sido propostos para o problema da extração de dados de documentos textuais. Na área de Bibliotecas Digitais, a extração automática de metadados bibliográficos é um campo que tem ganhado muita atenção recentemente. Em [Han et al. 2003] é descrito um método de extração de metadados presentes em cabeçalhos de artigos científicos baseado em Support Vector Machines (SVM), o qual supera outros métodos de aprendizagem automática na mesma tarefa. Em [Day et al. 2005] é proposta uma abordagem para extração de metadados baseada em conhecimento ontológico. Esta abordagem exige que uma ontologia seja construída manualmente.

Em [Peng and McCallum 2006], os autores abordam o problema da extração de metadados bibliográficos e propõem a utilização de *Conditional Random Fields (CRF)* para resolver este problema. CRF 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. Na seção de experimentos, apresentamos um estudo comparativo entre este

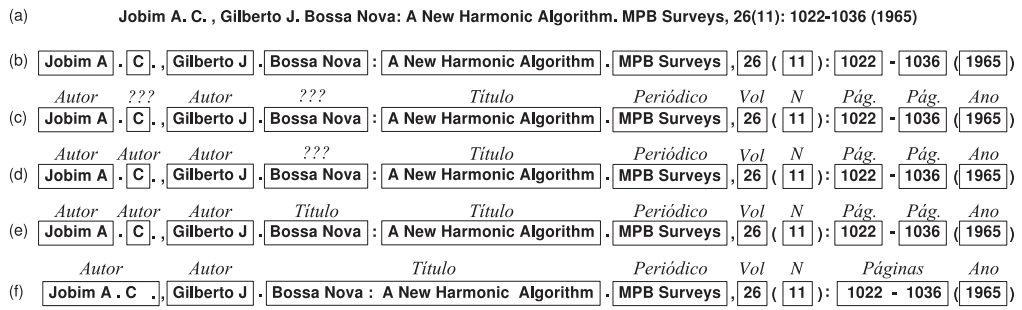

Figura 1. Exemplo de uma citação (a) e cada passo da extração: blocking (b), matching (c), binding (d, e), e joining (f).

método e a nossa abordagem.

3. O Método FLUX-CiM

Nesta seção, apresentamos nosso método de extração de metadados bibliográficos. Iniciamos apresentando alguns conceitos básicos e em seguida detalhamos cada uma das fases que compõem o método.

Conceitos Básicos

Em nosso trabalho, modelamos uma base de conhecimento (BC) como 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 *ocorrência*. Intuitivamente, O_i é o conjunto de valores típicos do campo m_i. Na Figura 2 é apresentado um exemplo simples de uma base de conhecimento, onde são ilustrados somente dois campos de metadados bibliográficos: *Autor* e *Título*.

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

Figura 2. Exemplo de uma Base de Conhecimento.

Outro conceito importante no âmbito deste trabalho é o de *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.

3.1. Fases do Método

3.1.1. Blocking

A primeira fase do nosso método de extração, consiste em dividir uma citação em *substrings* que chamamos de *blocos*. Em nosso método, consideramos cada bloco como um 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 1(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.

3.1.2. 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. Na fase de matching, valores textuais (por exemplo título) são manipulados utilizando uma função de similaridade que chamamos *FF*, tal função é definida a seguir:

$$FF(b, m_i) = \frac{\sum_{t \in T(m_i) \cap T(b)} \textit{fitness}(t, m_i)}{|T(b)|} \qquad fitness(t, m_i) = \frac{f(t, m_i)}{N(t)} \times \frac{f(t, m_i)}{f_{max}(m_i)} \quad (1)$$

onde a função FF estima a probabilidade de um bloco b fazer parte de uma ocorrência do campo bibliográfico m_i, através da avaliação de quão típicos os termos de b são em relação às ocorrências deste campo de acordo com a base de conhecimento. Para isso, utilizamos a medida de *fitness*, 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 frequência 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. 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, ano, volume, etc.) tradicionais funções de similaridade textual não funcionam corretamente. Assim, para atributos numéricos, 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:

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

onde σ e μ são o desvio padrão e a média, respectivamente, dos valores do campo bibliográfico m_i.

Após a fase de matching, a maioria dos blocos está associada 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 1(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.

3.1.3. Binding

A fase de binding associa os blocos unmatched restantes com campos bibliográficos. Existem três casos distintos que consideramos: *vizinhança homogênea*, *vizinhança parcial* e *vizinhança heterogênea*. Detalhamos abaixo a estratégia específica de binding que

foi adotada no caso de *vizinhança heterogênea*. Por questão de espaço foram omitidas neste texto os detalhes sobre o processo de binding para os casos de *vizinhança homogênea* e *vizinhança parcial*.

Vizinhança Heterogênea
Considere o exemplo na Figura 1(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. 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. Por exemplo, na Figura 1, 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 mais bem 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.

Considere a função $D(p_k, m_l, m_r)$ que estima a probabilidade de p_k ser um delimitador típico de campo entre os valores de m_l e 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 1(e), por exemplo, o bloco contendo o termo "Bossa Nova" é então associado ao campo *Título*, dado que $D(\text{“ : ”}, Titulo, Autor) < D(\text{“.”}, Titulo, Autor)$.

3.2. 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. Por exemplo, os blocos do campo *Autor* na Figura 1(e), devem ser unidos, para formar os valores de *Autor* como é ilustrado na Figura 1(f).

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 η. 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 (η).

4. Experimentos

Nesta seção, apresentamos uma comparação experimental entre o nosso método e o CRF, método considerado o estado da arte em extração de dados bibliográficos de artigos científicos. 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.

4.1. 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. Para os experimentos do domínio CS1, utilizamos uma coleção de citações da *PubMed Central (PMC)*[4]. No caso do domínio CS2, a coleção foi obtida a partir da Biblioteca Digital Scielo[5]. As coleções CS1 e CS2 são compostas por mais de 10 mil citações bibliográficas, destas, foram selecionadas 5 mil para compor a base de conhecimento e 2 mil citações para comporem a base de teste. Para a avaliação dos métodos de extração, utilizamos as medidas: precisão, revocação e Medida F [Cortez et al. 2009]. As configurações experimentais utilizadas na execução do método de extração CRF foram as mesma utilizadas no método FLUX-CiM.

5. Resultados

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 1. Resultados de medida F para as coleções CORA (a), CS1 (b) e CS2 (c).

[4] http://www.pubmedcentral.nih.gov/

[5] http://www.scielo.org/

Para todas as três coleções, executamos a implementação de CRF publicamente disponível [6]. Para assegurar uma comparação justa, o mesmo conjunto de registros de citações foi usado para treinar o 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.

Observando os resultados apresentados na Tabela 1, 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 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.

5.1. 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 experimentos em que os conjuntos de teste incluem citações com estilos distintos.

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 usada 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. Para construir a base de conhecimento para o FLUX-CiM e treinar o modelo do CRF escolhemos aleatoriamente 5.000 registros de citações em seu estilo original, ou seja, Estilo 1, de cada coleção respectiva.

Os resultados deste experimento são apresentados na Tabela 2. 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.

6. Conclusões

Neste artigo, apresentamos um novo método, FLUX-CiM, para extrair componentes (por exemplo: nomes de autor, títulos de artigos, locais, números de página) de citações

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

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

(a) CS1

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

(b) CS2

Tabela 2. Medida F obtida com diferentes estilos de citações para CS1 e CS2.

bibliográficas. Ao contrário dos métodos anteriores encontrados na literatura o nosso método não depende de padrões específicos para codificação de delimitadores. 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.

Realizamos 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. Como trabalho futuro, consideramos 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

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 ACM/IEEE JCDL 2007 Conference on Digital Libraries*, pages 215–224.

Cortez, E., da Silva, A., Gonçalves, M., Mesquita, F., and de Moura, E. (2008). FLUX-CIM: flexible unsupervised extraction of citation metadata. *Demo Section of Brazilian Symposium in Databases - SBBD*.

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):1144-1158, 2009.

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.

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.

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