Uma Proposta de Sistema de Busca para Recuperação de Formulários Digitais

Afonso Henrique Anastácio Calábria[1], Talles Brito Viana[1]

[1]Laboratory of Information System (LaIS)
Instituto Federal de Educação Ciência e Tecnologia do Ceará – Crato, CE – Brasil
ahacalabria@gmail.com, tallesbrito@ifce.edu.br

Abstract. *The larger quantity of generated documentation within organizations makes difficult to store, retrieve and track the processing of documents. In such a scenario, it is important to note a kind of document: the forms, in which these most often are created digitally, then are printed on paper and transmitted between people and departments in an organization according to an organizational process flow. This paper proposes a new approach for storage, indexing and retrieval of digital forms based on the definition of: user interfaces for creating digital forms, repositories of digital forms, search engines for indexing digital forms, and finally, user interfaces for query and retrieval of digital forms.*

Resumo. *A quantidade substancial dos documentos das organizações dificulta o armazenamento, recuperação, acompanhamento e processamento de documentos. Nesse cenário, é importante ressaltar um tipo de documento: os formulários, nos quais em sua maioria são criados digitalmente e então, impressos em papel e transmitidos entre pessoas e departamentos de uma organização seguindo um fluxo de processo. Este trabalho propõe uma nova abordagem para armazenamento, indexação e recuperação de formulários digitais baseada na definição de: interfaces para criação de formulários digitais, repositórios e motores de busca para indexação de formulários digitais e interfaces para pesquisa e recuperação de formulários digitais.*

1. Introdução

A produção de documentos e informações na sociedade atual é cada vez maior [Oliveira 2014]. Considerando que quantidade substancial dos documentos das organizações ainda é tratada em forma de papel, surge a seguinte problemática: Quão difícil é armazenar documentos na forma de papel? A grande quantidade de documentação gerada dentro das organizações, como por exemplo, formulários, memorandos, requisições e documentação fiscal, torna difícil a tarefa de armazenar, recuperar ou acompanhar o processamento dos documentos dentro da organização [Andrade 2002]. Especialmente destaca-se um tipo de documento: os formulários, estes que na maioria das vezes são criados digitalmente, em seguida são impressos em papel e transmitidos entre pessoas e departamentos. O grande problema do uso formulários impressos em papel está na dificuldade para entregar e repassar estes documentos, bem como, na dificuldade de recuperar e processar informações de formulários previamente criados e armazenados.

Nesse contexto, este trabalho propõe um Sistema de Busca para recuperação de documentos, especificamente, formulários digitais, com ênfase na descrição semiestruturada e indexação dos mesmos, com o objetivo de facilitar a gestão de formulários digitais. O Sistema de Busca proposto é detalhado na Seção 3.

2. Trabalhos Relacionados

No sistema proposto por [Barrus, Schwartz 2014] é demonstrado que o preenchimento de formulários através da captura de dados produzidos por informação manuscrita, via uma *stylus pen* em dispositivos móveis como um *tablet*, é uma evidência de que de fato o papel pode ser substituível de forma funcional e ecológica. No protótipo *Youfile* [Barchetti et al 2008], para que os documentos sejam indexados de maneira adequada é necessário que um operador humano associe manualmente os campos (ou partes) dos documentos com informações correspondentes aos termos e relações predefinidas através de ontologias. Apesar disso, o desempenho e a eficiência são dois fatores que em geral dificultam o emprego dessa técnica na prática. Em contrapartida, a proposta apresentada nesse trabalho opta por utilizar um algoritmo de recuperação de informação baseado na indexação de informações textuais, de forma a facilitar e automatizar o processo de indexação, tornando-o mais eficiente.

3. Solução Proposta

Para a descrição dos formulários digitais propõe-se o emprego de uma linguagem de descrição padronizada de formulários digitais, denominada XFDL (*Extensible Forms Description Language*) [Boyer et al 1998]. XFDL é uma linguagem que possibilita descrever e tratar a representação de documentos digitais do tipo formulário. Além disso, a XFDL permite descrever a estrutura de formulários digitais, visando lidar com problemas como: precisão de layout, validação, assinaturas digitais e suporte a transações [Boyer et al 1998]. Neste sentido, o sistema de busca proposto neste trabalho lida com armazenamento, indexação e recuperação de uma fonte de descrições XFLD, que englobam formulários digitais vazios (ainda não preenchidos), bem como, formulários já preenchidos com informações. Assim, é proposta uma nova abordagem para o armazenamento, indexação e recuperação de formulários digitais que é baseada na definição de um novo Sistema de Busca de formulários digitais com a arquitetura ilustrada na Figura 1.

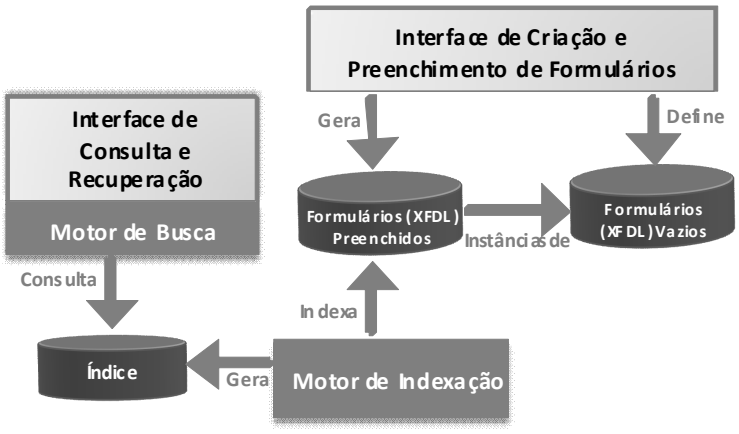

Figura 1. Arquitetura do sistema de busca proposto

A arquitetura proposta é composta dos módulos: *(i) Interface de Criação e Preenchimento de Formulários:* constitui uma interface gráfica que permite criar novas descrições XFDL de formulário vazios ou preenchidos e armazená-las em repositórios. *(ii) Repositórios de Formulários Digitais:* repositórios para o armazenamento de formulários XFDL vazios e preenchidos. *(iii) Motor de Indexação:* pode ser visto como o módulo responsável por indexar os formulários preenchidos descritos em XFDL com a finalidade de gerar índices específicos através do emprego de algum algoritmo para indexação de dados semiestruturados (por exemplo, um algoritmo clássico para esta finalidade é discutido em [Kotsakis 2002]). *(iv) Motor de Busca* e *Interface de Consulta e Recuperação:* mediante uma função de ranqueamento retorna os resultados mais relevantes a uma consulta de um usuário que procura por formulários digitais.

4. Considerações Finais

Por fim, é importante ressaltar que o escopo deste trabalho está em definir uma proposta de um novo sistema de busca de formulários digitais em um alto nível de abstração. Como passo direto de continuação deste trabalho têm-se os seguintes direcionamentos: primeiramente a definição detalhada e formalizada do algoritmo de indexação e ranqueamento que será adotado no sistema proposto. Em seguida, será desenvolvida uma implementação dos módulos da proposta apresentada neste trabalho. Em estudos preliminares foi concluído que, em termos de viabilidade tecnológica, a concretização das ideias abstratas apresentadas neste trabalho é plenamente possível. Para isto, será empregado o uso da API Lucene [McCandless et al 2010] (uma API escalável e de alto desempenho para indexação e recuperação de informação) na construção dos módulos dos Motores de Busca e Indexação. Assim, com uma futura concretização (implementação, experimentação e validação) do Sistema de Busca proposto, espera-se que o emprego deste vise reduzir na prática o uso de formulários de papel.

Referências

Andrade, M. (2002) “Gerenciamento Eletrônico da Informação: Ferramenta para a Gerência Eficiente dos Processos de Trabalho”. Seminário Nacional de Bibliotecas Universitárias, Recife.

Barchetti, Ugo et al. (2008) “How Can Ontologies Support Enterprise Digital And Paper Archives?: A Case Study”, In: Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. p. 627-636. ACM.

Barrus, John W. Schwartz, Edward L. (2014) “Image-Based Document Management: Aggregating Collections of Handwritten Forms”, In: Proceedings of the 2014 ACM symposium on Document engineering. ACM, p. 117-120.

Boyer, J., Bray, T., & Gordon, M. (1998) “Extensible Forms Description Language (XFDL) 4.0”, Draft Specification NOTE-XFDL-19980902, W3C. Acessado em Agosto de 2015: http://www.w3.org/TR/NOTE-XFDL.

Kotsakis, E. (2002) “Structured Information Retrieval in XML documents”. Proceedings of the 2002 ACM Symposium on Applied Computing (SAC'02). pp. 663-667.

McCandless, M., Hatcher, E., and Gospodnetic, O. (2010), Lucene in Action, Second Edition: Covers Apache Lucene 3.0, Manning Publications, Greenwich, CT, USA.

Oliveira, C.T. (2014) “O Gerenciamento Eletrônico de Documentos Sob a Ótica da Representação da Informação Arquivística”, Archeion Online 2.1.