An automatic linking service of document images reducing the effects of OCR errors with latent semantics

Renato F. Bulcão-Neto,
José Camacho-Guerrero
Innolution Sist. de Informática
Ribeirão Preto-SP, Brazil
renato.bulcao@acm.org
jcamacho_jr@yahoo.com

Álvaro Barreiro,
Javier Parapar
University of A Coruña
A Coruña, Spain
barreiro@udc.es
javierparapar@udc.es

Alessandra A. Macedo
Universidade de São Paulo
FFCLRP-DFM
Ribeirão Preto-SP, Brazil
ale.alaniz@usp.br

ABSTRACT

Robust Information Retrieval (IR) systems have been demanded due to the widespread and multipurpose use of document images, and the high number of document images repositories available nowadays. This paper presents a novel approach to support the automatic generation of relationships among document images by exploiting Latent Semantic Indexing (LSI) and Optical Character Recognition (OCR). The LinkDI service extracts and indexes document images content, obtains its latent semantics, and defines relationships among images as hyperlinks. LinkDI was experimented with document images repositories, and its performance was evaluated by comparing the quality of the relationships created among textual documents and among their respective document images. Results show the feasibility of LinkDI relating OCR output with high degradation.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Design, Experimentation, Measurement, Performance

1. INTRODUCTION

Document imaging is commonly used to describe software system that captures and stores images for later access. Recently there has been a tremendous increase in the number of document images repositories for multipurpose use including reading, teaching and research.

For instance, the Clendening Library publishes digital images from medical and natural history texts to stimulate the educational use of these images [15]. Historical newspapers editions (e.g. Today's News-Herald and New York Times) have been digitised by Google Corporation allowing users to automatically create timelines which show selected search results from relevant time periods [6]. Millions of patents and patent applications are indexed from the United States Patent and Trademark Office, allowing web users to search and scroll through pages, and zoom in on image areas [7].

However, the widespread use of large databases of document images has also demanded robust ways of images content indexing and retrieval. Latent Semantic Indexing (LSI) [5] is a technique of matrix processing widely adopted in Information Retrieval (IR) systems, which extracts the latent meaning of documents content from text-based document collections. LSI can intrinsically identifies relationships between words and the respective stem forms overcoming usual problems of lexical IR approaches (e.g. polysemy and synonymy). Authors advocate using the LSI capability of mitigating Optical Character Recognition (OCR) errors by manipulating the latent semantics of concepts in a document image retrieval system.

This paper presents the LinkDI (Linking of Document Images) service, which implements a novel approach to support the automatic generation of relationships among document images by exploiting both LSI and OCR. The LinkDI architecture, which extends previous work on linking services [10, 2], extracts content from document images using an OCR algorithm, indexes the textual information previously collected, processes the content indexed considering its latent semantics and redundancy, defines the relationships, and presents these as hyperlinks to users.

Experiments were carried out using two document images repositories automatically generated from textual documents by a ubiquitous capture and access system [12]. The first data set includes document images extracted from slides presentations of lectures on Computer Science, whereas the second data set describes document images extracted from slides presented by health care professionals in their meetings.

The objective with both experiments is to investigate whether relationships based on document images are as precise as relationships based on the respective textual documents.

In order to run the experiments, a public OCR algorithm with low recognition precision was used to simulate real situations from experimental context. Results show that LinkDI is effective in the case of relating OCR output with high degradation. Results also indicate that search engines could benefit from that approach to provide relationships among document images as in the previous use scenarios.

SAC'10 March 22-26, 2010, Sierre, Switzerland.

2. RELATED WORK

Research on how correctly retrieve information from degraded textual document collections has a long tradition in the IR field. Croft et al. [3] are pioneers on studies about the effectiveness of IR systems where results are based on actual OCR data. As a result of simulated OCR outputs on a variety of databases, they showed that low degraded OCR outputs have little effect on retrieval accuracy, but high degradation levels result in significant decrease of the retrieval effectiveness.

In order to prevent the loss of valuable information from presentations, Daddaoua et al. [4] also experimented with OCR and IR techniques. They compared the retrieval performance with slides transcriptions obtained through an OCR system over the slides' images against using a commercial software over the slides files. The performance using the OCRed transcripts was close to the performance using the text extracted by the commercial software.

In recent years, Magdy and Darwish [11] investigated the effect of OCR correction techniques on the effectiveness of retrieving Arabic document images using distinct index terms. Results show that effects on retrieval are recognisable only if the reduction of word error rates surpasses a given limit. Moreover, a very large language model for correction can reduce the need for morphologically sensitive error correction.

Related work has exploited the combination of OCR algorithms and IR systems to search and matching specific information, and document image indexing. Here, it is proposed the use of OCR as an extension of a linking infrastructure to automatically create relationships among text extracted from document images, using the latent semantics of document images content to mitigate the effects of OCR errors.

Furthermore, this LSI/OCR approach for document images is supported by the fact that statistical IR methods do not need perfectly clean data to work well. That is, a robust combination of IR techniques (e.g. LSI) and OCR-based conversion can represent a reasonable approach for the creation of relationships among document images.

3. THE LINKDI SERVICE

The LinkDigger infrastructure collects, analyses and interrelates textual document collections [10, 2]. The core of LinkDigger is the Latent Semantic Indexing (LSI), which is a statistical IR method capable of organising text objects into a semantic structure in which queries against these will return results that are conceptually similar in meaning to the query, even if they do not share a specific word or a set of words with the query.

Here the LinkDI service is presented as an extension of the LinkDigger infrastructure to generate relationships among document images content, which is extracted via OCR process. LinkDI then exploits LSI to deal with the misrecognition of characters problem of OCR-based document images.

3.1 LinkDI in details

The LinkDI service architecture is presented in Figure 1, and it is described as follows:

(1) LinkDI collects documents in multiple formats (HTML, XML, TIFF, BMP, JPEG, etc.) from remote and local URI's. In order to limit the amount of information to be collected, it is allowed to configure the depth level of URI's accessed. Compressed files are also collected and recursively extracted from ZIP and JAR formats.

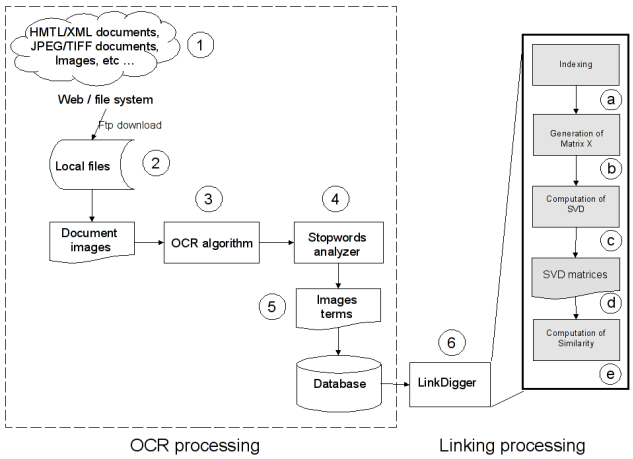

Figure 1: An overview of the LinkDI architecture.

(2) All documents collected are stored on a local directory. Textual documents are sent to the linking process (step 6), whereas document images follow the next step (3).

(3) LinkDI runs an OCR engine over images collected. Different OCR engines can be configured with LinkDI.

(4) Stopwords elimination, which reduces the set of index terms generated by the OCR process. LinkDI supports stopwords dictionaries for different languages (e.g. English, Portuguese and Spanish).

(5) All relevant words resulting from the stopwords elimination process are stored as an inverted index data structure called $ImagesTerms$.

(6) Information is related through the following process, which includes the LSI technique, among others:

- **(a) Indexing:** significant words (excluding stopwords) were extracted from documents. Besides, other text operations (e.g. identification of nouns groups) can be performed to facilitate documents representation with respect to index terms. Optional stemming may be carried out, what reduces each word to its linguistic root in accordance with the language detected. Currently, LinkDI supports documents in English, Spanish and Portuguese.
- **(b) Generation of a Term-Document Matrix:** a Term-Document matrix (or matrix X) is generated, where rows and columns represent index terms and documents, respectively. The importance of an index term is represented by a corresponding weight, which represents the term frequency (tf) combined with the frequency of the same term in the whole document collection (idf) [1]. Term weights are used by LinkDI to perform the degree of similarity between document pairs.
- **(c) Computation of SVD:** the matrix X is decomposed into three component matrices T, S and D' using Singular Value Decomposition (SVD),

which is part of the LSI technique. LSI has demonstrated to be extremely valid for this proposal because it feasibly manipulates "noise" produced by OCR misrecognition. The feasibility of using LSI is provided by enough redundancy information in matrix X compensating the amount of error caused by misrecognised characters.

The matrix S is a diagonal matrix with non-zero entries (called singular values) along a central diagonal. A large singular value indicates a large effect of this dimension on the sum-squared error of the approximation. By convention, S diagonal elements are constructed to be all positive and ordered in decreasing magnitude so that the first k largest singular values may be kept and the remaining smaller ones are discarded.

SVD has been extended to automatically select entries larger than 70% of sum of all entries of S. As a consequence, k-value considers a lost of 30% of information in LinkDI's current implementation. The tuning of k-value is widely investigated in the literature [5, 8].

(d) Manipulating SVD Matrices: the reduced dimensionality solution generates a vector of k real values to represent each document. SVD provides reduced rank-k approximation of a term-document matrix X for any value of k. The outcome is the reduced matrix $\hat{X}$ obtained by multiplying the three reduced component matrices.

(e) Computation of Similarity: in order to obtain the degree of similarity between documents given a particular index term, LinkDI's current version implements the cosine [13] technique over each pair of document vectors extracted from the matrix $\hat{X}$. As the angle between document vectors shortens, the cosine angle approaches 1, which means the highest similarity levels between document pairs. The outcome is a list of relationships between pairs of document images that can be used for search and recommendation purposes.

3.2 LinkDI in use: A case study

In previous work [9], authors argued for automatic linking to supporting authoring, extension and recommendation of text-based material before, during and after captured live experiences by capture and access (C&A) applications such as iClass [12]. In this work, LinkDI is a step further because it also supports authoring, extension and recommendation of captured material based on OCRed document images.

Before a lecture takes place, users upload textual documents (e.g. slides presentations) into iClass, which in turn converts those to corresponding document images to be presented during the lecture. As the lecture finishes, all information presented is automatically transformed into web-accessible hyperdocuments. For the purpose of this paper, authors focus on the C&A applications capability of producing images from textual documents (e.g. iClass) as well as the LinkDI capability of relating those document images.

In this case study, iClass captured, registered and documented not only lectures on Computer Science, but also quotidian events of different specialisations in Medicine such as symposiums. Before a symposium takes place, health care professionals upload Powerpoint-like slides with patient's clinical data into iClass repository including textual documents (i.e. clinical history) and image-based examinations. As the goal of this work is to compare the quality of links among pure text and among OCRed text, slides content is automatically converted to images by iClass, which are in turn collected by the OCR process of LinkDI. Afterwards, LinkDI performs all steps of the linking process as described in Figure 1.

During a symposium, slide images are presented on an electronic whiteboard upon which health care users may write using a digital pen or a mouse. iClass captures all users' interactions with slide images and registers them into an XML document representing the capture session — forward and backward slides navigation, digital ink-based handwritten notes (as foreground images), textual notes, etc.

As the symposium finishes, the XML document integrating all captured data is automatically transformed into hyperdocuments for different types of access. A web page with hyperlinks to those documents is also generated by iClass. The same capturing and automatic documentation process is carried out in the Computer Science lectures.

In order to access such documentation, user interfaces for LinkDI were developed so that users could be authenticated. Users are then allowed to access the syllabus web page, and a new web browser window is opened as a hyperlink is clicked on to show documentation content: (i) a web page with synchronised playback of slide images, handwritten notes and audio stream, or static web pages with (ii) titled slide images or (iii) slideshow navigation.

By navigating on a slide image, users may invoke LinkDI and request a list of slide images related to that. As the slide image reference was already collected and processed by LinkDI, this returns only the relationships also calculated in advance. Figure 2a depicts an example of slide image reference for LinkDI. As a result of LinkDI invocation, a new web browser window is opened and presents hyperlinks to slide images related to the image reference as in Figure 2b.

The next section describes experiments with LinkDI using sets of slide images generated by iClass from events of different knowledge areas.

4. EVALUATION AND RESULTS

The evaluation consisted of three different phases: (i) relationships generated among textual documents were evaluated against human experts; (ii) the same process was performed with relationships among document images; and (iii) both results of linking textual documents and linking the respective document images were also compared.

4.1 Test collections

The first experiment created relationships among lectures of Computer Science courses from three different university institutes. This test collection includes 205 lectures from 10 courses offered from 2004 to 2006 including Algorithms and Data Structures (55), Software Engineering (49), Object-Oriented Programming (29), Theory of Computation (26), among others (46). The relationships generated using this collection were analysed by a technical group with strong background in Computer Science.

The second experiment created relationships among meetings of health care professionals from a University Internal Medicine Department. The Medical Clinic (MC) collection

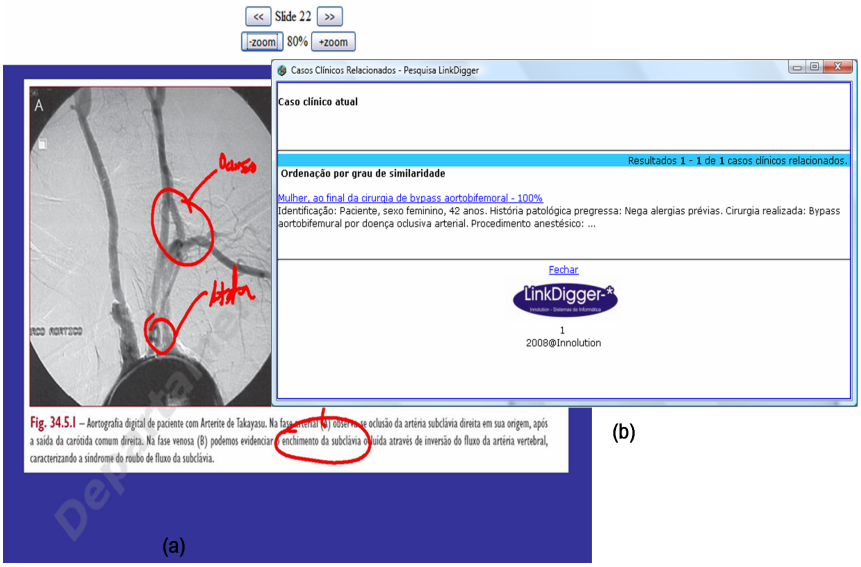

Figure 2: (a) document image reference; (b) hyperlink to a slide image of a clinical case related to (a).

contains 161 clinical cases of patients captured in medical grand rounds (51), scientific meetings (32), and symposiums of multiple specialisations in Medicine (78) from February to November of 2007. Medical grand rounds are weekly meetings in which a multidisciplinary team discusses clinical cases towards a diagnostic conclusion. The relationships generated using this collection were analysed by a multidisciplinary team of health care professionals.

Towards building IR reference collections, all possible relationships for each test collection were firstly produced. Then, members of evaluation teams laboriously classified each relationship through a high/medium/low/no relevance scale, which respectively represents 3/2/1/0 as reference value. Finally, the average of those assigned values for each relationship were computed, and only those with average value greater than 2 were considered relevant in this study.

4.2 Experiments and Results

After eliminating stopwords from the *CS text collection*, 66,640 words were collected and represented 13,006 terms. Once running LSI, LinkDI built a matrix X with 13,006 rows (terms) and 205 columns (lectures). In order to process matrix X with a 30% data loss rate, LinkDI computed the value of k equals to 87, i.e. only the 87 highest values in the singular matrix S were selected for the next steps.

Regarding the *MC text collection*, 14,724 words were collected after stopwords elimination, which included 10,852 terms. Similarly, LinkDI generated a matrix X with 10,852 rows and 161 columns (clinical cases), and it computed the value of k equals to 64, i.e. only the 64 highest values in the singular matrix S were take into account for the next steps.

In comparison with results of both *text collections*, 40% more terms were OCR recognised from *document image collections*. However, authors could observe that OCR processing added several misrecognised words as new terms, and only 3,839 terms were in both CS collections (text and image), for instance. That is an interesting scenario because it is featured by the use of degraded text by LinkDI, what may allow us to check whether the combined use of LSI (semantics) and OCR (misrecognition) results in an effective performance towards linking of document images.

Figure 3 depicts the results from both experiments. For both collections results, the x-axis describes the similarity threshold (i.e. cosine) used to filter the number of relationships (or hyperlinks) created from matrix $\hat{X}$, whereas the y-axis presents the values of weighted harmonic mean of precision and recall (also called F-measure) [14].

Similarity thresholds range from 10 to 90, i.e. hyperlinks with cosine greater than or equal to 0.1 and 0.9, respectively. From the CS collection, 12,322 links were retrieved with threshold value of 10, and only 80 hyperlinks connect documents with threshold value of 90.

As F-measure reaches the highest value, LinkDI retrieves the best precision value for the highest fraction of known relevant hyperlinks from the whole collection. For the MC collection, the best value of F-measure is 0.55 when cosine is 0.7. On the other hand, the worst value of F-measure suggests the worst precision recall ratio.

Curves in Figure 3 have similar increasing and decreasing behaviour of F-measure values. As the cosine threshold is closer to zero, most non-relevant hyperlinks are created. As the cosine threshold distances itself from zero, both curves grow up until a point where they start decreasing. Those knee points in Figure 3 represent the best performance of the LinkDI service for each collection.

It is important to explain the difference between text- and image-related curves in both graphics. Those differences were already expected because OCR usage implies a degradation of information if compared to the original information. Authors tried to reduce the impact of that problem using redundancy of information by means of LSI processing. The CS collection illustrates better that difference for cosine values between 0.4 and 0.8.

Another important point is that the quality of CS document images is not as good as document images of the MC collection. During the capture of lectures of the CS collection, iClass resized image slides to improve their visualisation on the Web. However, iClass started capturing and

Figure 3: X-axis describes cosine values, and y-axis represents harmonic means between precision and recall.

storing document images in their original size from 2007.

For that reason, authors believe that MC collection results have less degradation of information when LinkDI performed the OCR process. As a consequence, the distance between text and image-related curves in the MC collection is smaller. Due to reduction in effects of OCR errors with latent semantics, even using a non-accurate OCR algorithm, authors advocate the feasibility of the LinkDI service to automatically link document images.

5. CONCLUDING REMARKS

This paper presented the LinkDI (Linking of Document Images) service, which provides users with a novel use of OCR and IR technologies aiming to generate LSI-based hyperlinks among document images.

LinkDI could benefit scenarios where the goal is to retrieve and recommend information not only available as text, but also as document images. It could be useful for patent analysis by discovering relationships between a patent document reference and patent images. A few number of relationships returned may suggest an innovative product or methodology, or even trends in developing new technologies.

Experiments results indicate a positive direction of the proposal: a small distance between experimental curves with respect to document images and pure text only was obtained. From both experiments, linking document images is almost so precise as linking the respective textual information, what suggests the feasibility of manipulating OCR-based images with traditional text IR techniques towards automatic linking of document images.

6. ACKNOWLEDGEMENTS

We thank CNPq (557976/2008-1), FAPESP (03/07968-9, 04/12477-7, 05/60729-8, 05/60038-5, 06/58984-2), Ministerio de Ciencia e Innovación (TIN2008-06566-C04-04), FEDER, and Xunta de Galicia (07SIN005206PR) for the funding support.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] J. A. Camacho-Guerrero, A. A. Macedo, and M. G. C. Pimentel. A look at some issues during textual linking of homogeneous Web repositories. In *ACM Symposium on Document Engineering*, pages 74–83, USA, 2004.

[3] W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Symposium on Document Analysis and Information Retrieval*, pages 115–126, USA, 1994.

[4] N. Daddaoua, J. M. Odobez, and A. Vinciarelli. OCR based slide retrieval. In *International Conference on Document Analysis and Recognition*, pages 945–949, USA, 2005.

[5] S. Deerwester, S. T. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391 – 407, 1990.

[6] Google Corp. Google News Archive Search Homepage, 2009. `http://news.google.com/archivesearch`.

[7] Google Corp. Google Patent Search Homepage, 2009. `http://www.google.com/patents`.

[8] T. Jaber, A. Amira, and P. Milligan. Empirical study of a novel approach to LSI for text categorisation. In *IEEE Symposium on Signal Processing and Its Applications*, pages 1–4, UAE, 2007.

[9] A. A. Macedo, L. A. Baldochi Jr, J. A. Camacho-Guerrero, R. G. Cattelan, and M. G. C. Pimentel. Automatically linking live experiences captured through a ubiquitous infrastructure. *Multimedia Tools and Applications*, 37(2):93–115, 2008.

[10] A. A. Macedo, M. G. C. Pimentel, and J. A. Camacho-Guerrero. An infrastructure for open latent semantic linking. In *ACM Conference on Hypertext and Hypermedia*, pages 107–116, USA, 2002.

[11] W. Magdy and K. Darwish. Effect of OCR error correction on Arabic retrieval. *Information Retrieval*, 11(5):405–425, October 2008.

[12] M. G. C. Pimentel, L. A. Baldochi Jr, and R. G. Catellan. Prototyping applications to document human experiences. *IEEE Pervasive Computing*, 2(6):93–100, 2007.

[13] G. Salton and M. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

[14] W. M. Shaw, R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing Management*, 1(33):15–36, 1997.

[15] The Clendening Library Group. Digital Clendening, 2009. `http://clendening.kumc.edu`.