Journal of Universal Computer Science, vol. 17, no. 1 (2011), 64-80
submitted: 8/4/10, accepted: 12/12/10, appeared: 1/1/11 © J.UCS

The Use of Latent Semantic Indexing to Mitigate OCR Effects of Related Document Images

Renato F. Bulcão-Neto, José A. Camacho-Guerrero, Márcio Dutra
(Innolution Sistemas de Informática Ltda.
Ribeirão Preto-SP, Brazil
renato.bulcao@gmail.com, jcamacho_jr@yahoo.com, mdutra@gmail.com)

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

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

Abstract: Due to both the widespread and multipurpose use of document images and the current availability of a high number of document images repositories, robust information retrieval mechanisms and systems have been increasingly demanded. This paper presents an approach to support the automatic generation of relationships among document images by exploiting Latent Semantic Indexing (LSI) and Optical Character Recognition (OCR). We developed the LinkDI (**Link**ing of **D**ocument **I**mages) service, which extracts and indexes document images content, computes 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 as well as among their respective document images. Considering those same document images, we ran further experiments in order to compare the performance of LinkDI when it exploits or not the LSI technique. Experimental results showed that LSI can mitigate the effects of usual OCR misrecognition, which reinforces the feasibility of LinkDI relating OCR output with high degradation.

Key Words: Applied Computing, Information Retrieval, Document Engineering, Latent Semantic, Optical Character Recognition, Document Image, Experimentation.

Category: H.3, H.3.3, H.5.4

1 Introduction

Document imaging has been used to describe software systems that capture and store images for later access. Such documents are often scanned (or filmed) for archiving, and stored as images. Recently there has been a tremendous increase in the number of document images repositories for multipurpose use including reading, teaching and research.

For instance, historical newspapers editions (e.g. Today's News-Herald and New York Times) have been digitized by Google Corporation allowing users to automatically create timelines which show selected search results from relevant time periods [Google Corp., 2009a]. In another Google initiative, 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 [Google Corp., 2009b].

In order to support crossing of historical information, Brazilian government has digitized thousands of criminal records produced during Brazilian dictatorship [Proin, 2009]. A consortium of libraries from the University of São Paulo have also digitized rare literature to widespread its content [Obras Raras, 2009].

Document images may also have a great value for attorneys in case law — those are often made available as historical archives of print law reports. Old technical reports, dissertations and thesis have been digitized by most universities to make them searchable on the Web, mainly for research purposes. Hospitals have also struggled to digitize old health records with great contribution for surgical decision taking and epidemiology and clinical studies [Andreas et al., 2005].

Despite the widespread use of large databases of document images, traditional search engines usually exhibit unsatisfactory results from such documents content [Callan et al., 2002]. In that context, Optical Character Recognition (OCR) systems translate images of handwritten or typewritten text into machine-editable text, or translate pictures of characters into a standard encoding scheme representing them. One of the OCR's more challenging goal is to solve misrecognition of characters translations [Taghva et al., 1996], which causes a strong demand for robust ways of document images indexing and retrieval.

Latent Semantic Indexing (LSI) [Furnas et al., 1988] is a well-established technique of matrix processing widely adopted in Information Retrieval (IR) systems. By extracting the salient meaning of documents content from collections of textual documents, LSI overcomes usual problems of lexical IR approaches such as polysemy and synonymy exploiting redundancy in full text [Taghva et al., 1996].

From the light of document image retrieval, we developed the LinkDI (**Link**ing of **D**ocument **I**mages) service, which implements automatic generation of semantic relationships among document images by exploiting LSI and OCR together [Bulcão-Neto et al., 2010]. LinkDI extracts document images content through an OCR process, indexes and processes that content considering its latent semantics and redundancy, defines relationships, and presents these as hyperlinks to users.

We carried out experiments with document images repositories aiming to verify whether relationships based on document images are as precise as relationships based on corresponding textual documents. Results showed the feasibility of LinkDI relating OCR output with high degradation [Bulcão-Neto et al., 2010].

In this paper, we aim to reinforce the LSI capability of mitigating OCR misrecognition by manipulating the latent semantics of concepts in a document image retrieval system as LinkDI. In order to achieve this, we ran further experiments using LinkDI and the previous document images in two different situations: the creation of relationships among document images content with and without LSI processing. Results confirm that LSI can improve accuracy of textual manipulation after OCR processing by means of redundancy of information.

This paper is organized as follows. Section 2 reviews related work. Section 3 presents the LinkDI service. Section 4 details the experiments carried out and their respective results. Section 5 presents final remarks and future work.

2 Related Work

The study of how correctly retrieve information from degraded text collections has a long tradition in the information retrieval field [Beitzel et al., 2003]. As previously pointed out, the main problem when dealing with degraded information and traditional retrieval models has been the fail of term matching because of the degradation generated by OCR methods.

The first problem when developing new models to deal with degraded information was the absence of suitable collections. In order to tackle this problem and due to the difficulty of building test collections to obtain quantitative data on the impact of OCR errors on the accuracy of a text retrieval system, authors in [Croft et al., 1994] carried out evaluations using simulated OCR output on a variety of databases. Results showed that high quality OCR devices have little effect on the accuracy of retrieval, but low quality devices used with databases of short documents can result in significant degradation. In other words, the most important types of errors would not be random, but rather when a relevant document is made non-retrievable by poor quality scanning or word misrecognition.

In 1996 the TREC initiative built a collection to evaluate the effectiveness of retrieval systems when dealing with degraded information in the context of the Confusion Track [Kantor and Voorhees, 1996]. However, this collection was not useful to evaluate our LSI/OCR approach because the relevance judgements from Confusion Track included solely one document for query, and no other relevance judgements were also provided.

In order to address the loss of potentially valuable information from presentations, authors in [Daddaoua et al., 2005] applied IR technologies to speakers' slides used in real-world presentations. Slides transcriptions were obtained by using an OCR system over slides images. The same retrieval tasks were performed over transcriptions of the same slide corpus using a commercial software, which only extracted text from slides presentation files, or different versions of an OCR system. Experiments showed that the retrieval performance obtained using the OCR transcripts were close to the performance obtained with the text extracted using the commercial software. It was also shown that the OCR-based system had the ability to extract, index, and retrieve text embedded in figures (e.g. plots and diagrams) that were not often accessible to the commercial software.

In recent years, authors in [Magdy and Darwish, 2008] investigated the effects of OCR correction techniques on the effectiveness of retrieving Arabic document images using distinct index terms. Results showed that effects on retrieval are recognizable 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.

The literature has also reported efforts to detect and classify common noises in OCR'ed document images such as bleed through, framing, skew, orientation, and blur [Lins et al., 2010]. We believe that an OCR process preceded by such image filtering techniques might produce better results in terms of document images retrieval, e.g. if integrated with the LinkDI service.

OCR-free methods have also gained special attention in the past few years for various document image retrieval applications. Authors in [Tan et al., 2002] proposed a method for text retrieval from document images in which these are segmented into character objects, whose image features generate n-gram document vectors. Then, the text similarity between documents is measured by calculating the dot product of the n-gram document vectors. The n-grams are more promising than words due to the problem of segmentation into words in some languages like Chinese, and due to the language independence nature of the n-grams. The importance of the n-gram based approach has also been revisited in [Parapar et al., 2009] in which a distributed retrieval model based on multiple n-gram indexing achieved significant improvements in effectiveness. An OCR-free method for word spotting in document images was proposed in [Rios et al., 2010], which provides a significant reduction of processing time in the document segmentation process, and robustness to noisy documents.

In general, related work has exploited the combination of OCR algorithms and IR systems to search and matching specific information as well as to index document images content. Here we propose the use of an OCR process as an extension of a linking infrastructure towards automatically creating relationships among text extracted from document images. Such relationships are enriched by the LSI-based latent semantics of document images content, which we advocate as a feasible solution to mitigate the effects of OCR misrecognition.

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 fine. In other words, a robust combination of IR techniques (e.g. LSI) and OCR-based conversion might represent a reasonable approach for the creation of semantic relationships among document images content.

3 The LinkDI service

In the past few years we have investigated mechanisms allowing the automatic identification of relationships in textual homogeneous web-based repositories. As a result, we have built the LinkDigger software infrastructure, in which textual repositories can be analyzed and related [Camacho-Guerrero et al., 2004].

3.1 The linking infrastructure

First, we designed LinkDigger to manage an open linkbase to store relationships [Macedo et al., 2002b] and to allow user feedback over the relationships generated [Macedo et al., 2002a]. LinkDigger was also developed as the main component of a web page recommender service, called WebMemex [Macedo et al., 2003].

LinkDigger was part of an infrastructure, called CALiSP: Infrastructure for Capturing, Accessing, Linking, Storing and Presenting information, which automatically relates live experiences captured (e.g. lectures), considering the extraction of textual information as a background process [Macedo et al., 2008].

In previous work [Macedo et al., 2004], we argued that automatic linking services should be available as authoring support tools before, during and after live sessions supported by ubiquitous capture and access applications. We proposed that links generated using our proposal should be provided to users facilitating searching and recognition of desired information.

In short, all LinkDigger's challenges have exploited the LSI technique to organize text objects into a semantic structure appropriate for linking to overcome usual problems of lexical approaches such as polysemy and synonymy. In this paper, we aim to demonstrate the influence of LSI to solve the problem of misrecognition of characters of OCR'ed document images.

3.2 LinkDI in details

We present the LinkDI service developed from the LinkDigger infrastructure towards the automatic generation of relationships (as hyperlinks) using document images content extracted via OCR processing. The LinkDI underlying processing is presented in Figure 1.

(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.

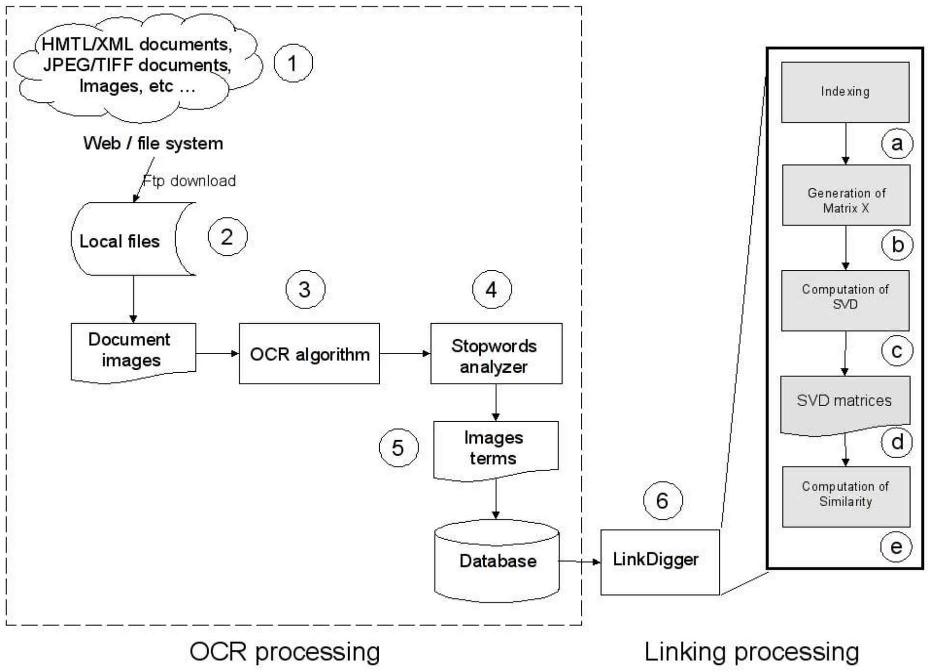

Figure 1: An overview of our LinkDI infrastructure for the automatic generation of relationships considering document images content.

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

(3) LinkDI infrastructure runs an OCR process adapted for its purposes, which converts document images back to text files [GOCR Group, 2008].

(4) In computing systems, textual documents can be represented as vectors, graphs and trees. With large collections, information retrieval systems might have to reduce the set of index terms. Words as conjunctions, prepositions and pronouns, named as stopwords, are eliminated, since they occur frequently in the text and they are words that do not represent the documents they belong to from a semantic perspective. Aiming to eliminate stopwords, stopwords dictionaries for different languages are collaboratively defined and distributed. So our optical character recognition processing generates a list of words extracted from image files which are analyzed by a *Stopword_Analyzer* to eliminate stopwords.

(5) A special data structure called $Words$ stores the relevant words extracted from the OCR process, the so-called images terms.

(6) Afterwards, terms are stored to be related through the following process:

(a) Indexing: significant words (excluding stopwords) were extracted from documents. Besides, other text operations can be applied to facilitate the representation of documents and allow the logical view from that of a full text to that of a set of index terms. So optional stemming is carried out, what reduces each word to its linguistic root based on the language detected. LinkDI's current implementation supports documents in English, Spanish and Portuguese. Other text operations can be assigned.

(b) Generation of Term-Document Matrix: the Term-Document matrix (or matrix X) is generated, where rows and columns represent index terms and documents, respectively. Not all terms are equally useful to represent document contents: less frequent terms allow identifying a narrower set of documents. The importance of the index terms is represented by weights associated to them. Every term is given an appropriate weight as its term frequency (tf) combined with the frequency of the same term in the whole documents collection, which is the inverse document frequency (idf) [Baeza-Yates and Ribeiro-Neto, 1999]. These term weights are used by LinkDI to compute a degree of similarity between each pair of documents.

(c) Computation of SVD: the matrix X is decomposed into three component matrices T, S and D' using Single Value Decomposition (SVD), which is part of the LSI technique [Furnas et al., 1988].

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 discarded. We extended SVD to automatically select entries larger than 70% of sum of all entries of S. Consequently our value of k considers a lost of 30% of information.

(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 $\hat{X}$, which is obtained by multiplying the three reduced component matrices.

(e) Computation of Similarity: in order to obtain the similarity level between documents given a particular term, LinkDI's current version implements the cosine [Salton and Lesk, 1968] 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 each pair of document images that can be used for search and recommendation purposes.

The OCR process of the LinkDI infrastructure, which is depicted in the left hand of Figure 1, can be further described in [Camacho-Guerrero et al., 2007].

3.3 LinkDI in use: A case study

In previous work [Macedo et al., 2008], we argued for automatic linking to support authoring, extension and recommendation of textual material before, during and after captured live experiences by capture and access (C&A) applications such as iClass [Pimentel et al., 2007]. In this work, LinkDI is a step further because it also supports authoring, extension and recommendation of captured material based on OCR'ed 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 specializations in Medicine 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 main goal of this work is to compare the quality of links among OCR'ed document images with and without LSI, 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 — slides navigation, digital ink-based handwritten notes (as foreground images), textual notes, etc.

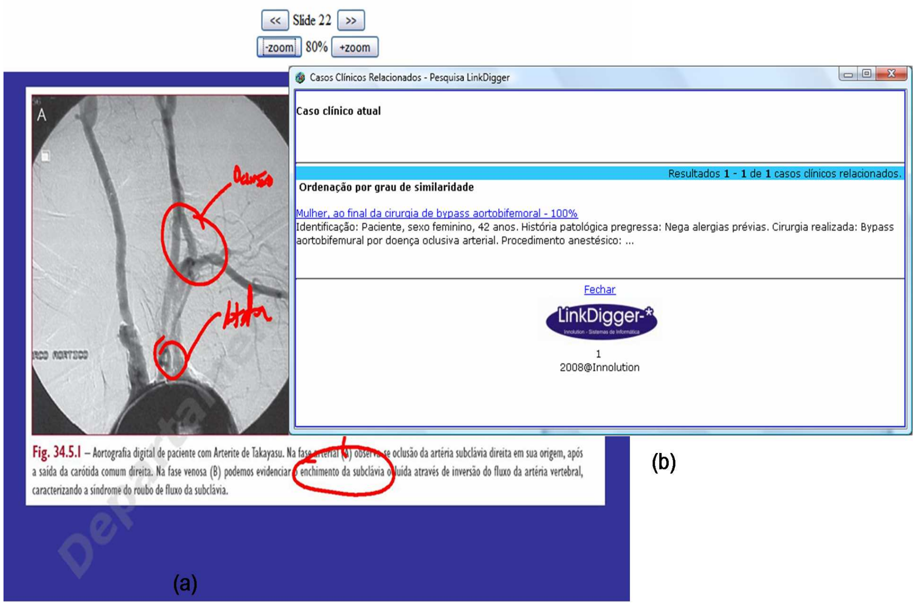

Figure 2: (a) Document image reference; (b) Hyperlink to a slide image of a clinical case related to (a).

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 automatic capture and documentation process is carried out for the Computer Science lectures.

In order to access such documentation, LinkDI user interfaces 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 synchronized 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 the current slide. As the slide image reference was already collected and processed, LinkDI returns only the relationships computed 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 with 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

First of all, our experiments tried to evaluate if links generated considering text extracted from document images are as precise as results extracted from pure textual documents. The first set of evaluation consisted of three different phases:

1. relationships generated among textual documents were evaluated against human experts;
2. the same process was performed with relationships among document images;
3. both results of linking textual documents and linking the respective document images were also compared.

Afterwards, we ran further experiments to compare relationships resulting from (i) linking of document images considering LSI and (ii) linking of document images not using LSI. We believed that LSI-based latent semantics reduced the effects of OCR errors.

4.1 Test Collections

Experiments created relationships among lectures of Computer Science courses from three different institutions. The Computer Science (CS) collection includes 205 lectures of 10 courses offered from 2004 to 2006 as described in Table 1.

Table 1: Computer Science (CS) collection

Courses	Lectures
Software Engineering	49
Object-Oriented Programming	29
Algorithms and Data Structures I	30
Algorithms and Data Structures II	25
Human-Computer Interaction	22
Parallel Computing	7
Computer Networks	9
Operational Systems	6
Computational Theory	26
Programming Computer	2
	205

Other set of experiments created relationships among meetings of health care professionals from a University Internal Medicine Department. The Medical Clinic (MC) collection contains 161 clinical cases of patients discussed from

February to November of 2007, as presented in Table 2. Clinical meetings include medical grand rounds, scientific meetings and symposiums of different specializations in Medicine. Medical grand rounds are weekly meetings in which a multidisciplinary team discusses clinical cases towards a diagnostic conclusion. Scientific meetings support scientific collaboration carried out by groups of researchers. Symposiums are thematic sporadic events with participation of external speakers.

Table 2: Medical Clinics (MC) collection

Events	Clinical cases
Medical grand round	51
Scientific Meeting	32
Symposium	78
	161

Towards building IR reference collections, all possible relationships for each test collection were firstly produced. Then, members of evaluation teams with strong background in their respective areas 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 Metrics

All experiments were evaluated in terms of the classical information retrieval measures recall and precision. The former is the fraction of known relevant links which were effectively extracted (completeness), whereas the latter is the fraction of retrieved links which are known to be relevant (fidelity or exactness).

In order to evenly weight precision and recall, i.e., good precision with reasonable recall, we calculated the harmonic mean F-measure (see Equation 1). F-measure is a weighted harmonic mean of precision and recall [Shaw et al., 1997].

$$F = 2/((1/precision) + (1/recall)) \tag{1}$$

The main chosen metric is the harmonic mean F-measure [Shaw et al., 1997], which assumes a high value only when both recall and precision are high.

4.3 Experiments and Discussion

After eliminating stopwords from the *CS textual collection*, 66,640 words were collected and represented 13,006 terms (distinct words). Once running LSI on this collection, LinkDI built a matrix X with 13,006 rows (terms) and 205 columns (lectures). In order to process matrix X with a 30% of data loss rate, LinkDI computed the value k equals to 95, i.e. only the 95 highest values in the singular matrix S were selected for the next steps.

Regarding the *MC textual 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 taken into account for the next steps.

After OCR process, 50% more terms (on average) were OCR recognized from both *document image collections*. For example, we obtained 18,170 OCR'ed distinct words without stopwords from *CS document images collection*, which represents 40% more terms than the *CS textual collection*. Authors could observe that OCR process added several misrecognized words as new terms, and only 3,839 terms were in both CS collections (text and image), for instance. The words which are not in both collections can not be discarded because the textual collections are not dictionaries from a specific domain vocabulary. For instance, lecture_ids and diseases are ordinary terms composed of letters and codes which can be equivocally captured by the OCR process generating new terms. As a result, many OCRed terms may lead to a worst performance. Our experimental scenario is interesting 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 results from two experiments: (i) linking considering text collections, and (ii) linking considering document image collections. For both graphics, 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 F-measure.

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 are 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 behavior of F-

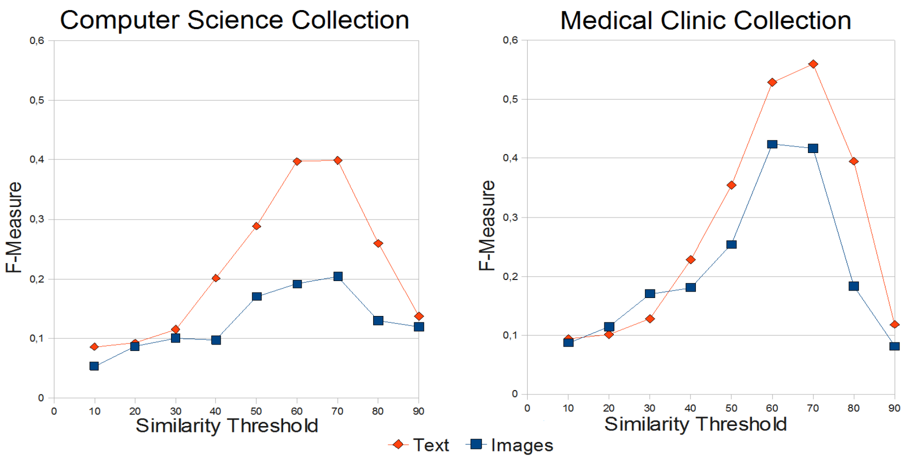

Figure 3: Graphics from experimental results. X-axis describes similarity thresholds (cosines) and y-axis represents harmonic means between precision and recall (F-measures).

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. The CS collection better illustrates that difference for cosine values ranging between 0.4 and 0.8.

An important point is that the quality of *CS document images collection* 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 visualization on the Web. However, iClass started capturing and 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.

Authors believed that redundancy of information by means of LSI processing could be able to reduce the impact of OCR degradation. We ran further experiments to verify this premise. Two experiments created relationships considering only the *CS and MC document images collections*. The first experiment carried out LSI processing before linking the OCR'ed information from document

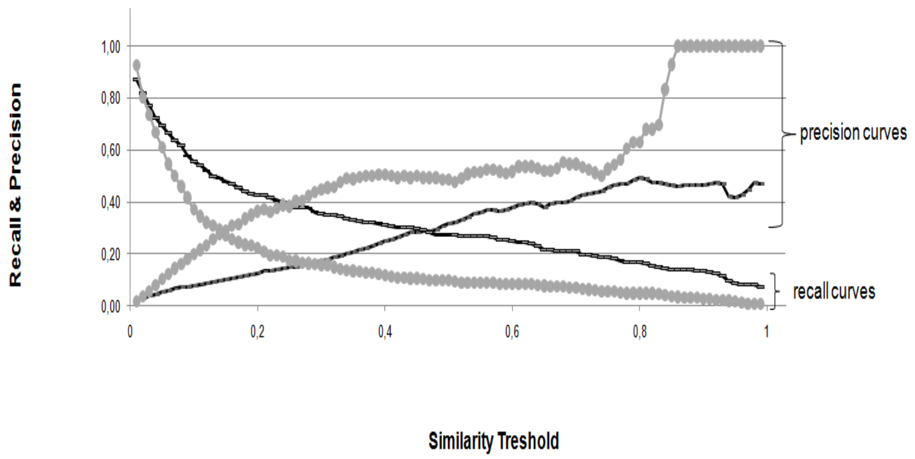

Figure 4: Graphics from experimental results. X-axis describes similarity thresholds (cosines) and y-axis represents precision and recall.

images. The second experiment relates the same information, but without LSI processing. Figure 4 presents results from both experiments. The x-axis depicts the similarity threshold (cosine), whereas the y-axis presents the values of precision and recall. Gray lines represent experiments without LSI and black lines are results from experiments with LSI process.

In comparison with results from precision curves, experiments without LSI present better values of precision. Otherwise, the recall curves depict opposite behavior. Experiments with LSI presented more known relevant hyperlinks effectively extracted (20% more hyperlinks than the experiments without LSI). This effect was already expected once redundancy of information by means of LSI would increase the completeness of relevant hyperlinks.

The best point (higher values of both recall and precision) for experiments without LSI occurs when cosine threshold is close to 15%; as a consequence, several (300) hyperlinks were defined by LinkDI from which 102 are relevant. On the other hand, the best point for experiments with LSI is close to cosine equals 50% when more relevant hyperlinks (92) are suggested to users from 252 hyperlinks created by LinkDI. For example, experiments without LSI generated 33 precise hyperlinks when cosine equals to 50%. Considering the same value of cosine, experiments with LSI created 93 relevant hyperlinks.

The experimentation comparing results with and without LSI suggests this could improve accuracy of textual manipulation after OCR process by means of redundancy of information. 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 Conclusions and Future Work

This paper presented the LinkDI service, which combines OCR and IR technologies to automatically generate relationships among document images. The main contribution of LinkDI is the computation of the semantics of document images content after an OCR process by means of the LSI technique, which distinguishes terms that better represent documents semantic. According to our experiments, LSI proved to be feasible because it can fairly handle noise produced by OCR misrecognition. This feasibility is provided by enough redundancy information in matrix X, which compensates the number of errors caused by misrecognized characters.

The latent semantics of the relationships computed by LinkDI could benefit scenarios where the goal is to retrieve and recommend information not only available as text, but also as document images. For instance, LinkDI could be useful for patent analysis by discovering relationships between a patent document reference and patent images, or even for the retrieval of OCR-based versions of the Cranfield reports, the Salton ISR series, and other old publications out of print, which have been digitized by authors in [Harman and Hiemstra, 2008].

Regarding the performance of the LinkDI linking infrastructure, the most challenging limitation is the O(mnc) complexity of the LSI method, where m is the number of documents, n is the number of terms, and c is the number of non zero elements in matrix X.

Experimental results indicate a positive direction of our proposal. Former results show a small distance between experimental curves with respect to document images and pure text only was obtained. Linking document images is almost so precise as linking the respective textual information. Later results suggest LSI is able to improve results from textual manipulation of OCR'ed documents. Experiments with LSI and document images resulted in better recall and F-measures values, and almost the same precision values. As a result, our experiments suggest the feasibility of manipulating OCR-based images with traditional text IR techniques towards automatic linking of document images.

Currently, we have set up experiments to evaluate this technique considering another realistic context such as an attorneys' office. As future work we intend to investigate other OCR engines with different precision levels. Our proposal can also be applied to other context, for instance, audio speech recognition (ASR) output because this suffers from the same problems of OCR processing in terms of misrecognized information. For this, there is a version of the iClass application being used in experiments that performs ASR over captured users' voices.

Acknowledgements

We thank CNPq (557976/2008-1), FAPESP (05/60038-5, 05/60729-8, 06/58984-2, 09/14292-8, 2009/05504-1), Ministerio de Ciencia e Innovación (TIN2008-06566-C04-04), FEDER, and Xunta de Galicia (07SIN005206PR) for the funding support. Authors also would like to thank and congratulate Innolution Sistemas de Informática for supporting scientific research.

References

[Andreas et al., 2005] Andreas, H., Klaus, S., and Matthias, W. (2005). Ubiquitous computing for hospital applications: RFID-Applications to enable research in real-life environments. In *Proceedings of the 29th Annual International Computer Software and Applications Conference*, pages 19–20. IEEE.

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

[Beitzel et al., 2003] Beitzel, S. M., Jensen, E. C., and Grossman, D. A. (2003). Retrieving OCR text: A survey of current approaches. In *Symposium on Document Image Understanding Technologies*, USA.

[Bulcão-Neto et al., 2010] Bulcão-Neto, R. F., Camacho-Guerrero, J., Barreiro, A., Parapar, J., and Macedo, A. A. (2010). An automatic linking service of document images reducing the effects of ocr errors with latent semantics. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 13–17, New York, NY, USA. ACM.

[Callan et al., 2002] Callan, J., Kantor, P., and Grossman, D. (2002). Information Retrieval and OCR: From converting content to grasping meaning. *SIGIR Forum*, 36(2):58–61.

[Camacho-Guerrero et al., 2007] Camacho-Guerrero, J., Bulcão-Neto, R., Carvalho, A., and Macedo, A. (2007). OCR processing to semantically linking image content. In *XIII Simpósio Brasileiro em Sistemas Multimídia e Web*, pages 25–27, Brazil.

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

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

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

[Furnas et al., 1988] Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., and Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *Intl. Conference on Research & Development in Information Retrieval*, pages 465–480.

[GOCR Group, 2008] GOCR Group (2008). Open-Source Character Recognition. `http://jocr.sourceforge.net`.

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

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

[Harman and Hiemstra, 2008] Harman, D. and Hiemstra, D. (2008). Saving and accessing the old ir literature. *SIGIR Forum*, 42(2):16–21.

[Kantor and Voorhees, 1996] Kantor, P. B. and Voorhees, E. M. (1996). Report on the TREC-5 confusion track. In *TREC*.

[Lins et al., 2010] Lins, R. D., Banergee, S., and Thielo, M. (2010). Automatically detecting and classifying noises in document images. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 33–39, New York, NY, USA. ACM.

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

[Macedo et al., 2004] Macedo, A. A., Camacho-Guerrero, J. A., Cattelan, R. G., Inacio Jr, V. R., and Pimentel, M. G. C. (2004). Interaction alternatives for linking everyday presentations. In *Proceedings of the ACM Conference on Hypertext and Hypermedia*, pages 112–113.

[Macedo et al., 2002a] Macedo, A. A., Camacho-Guerrero, J. A., and Pimentel, M. G. C. (2002a). Incluindo abordagens de recuperação de informação em serviços de criação de hiperligações. In *XXVIII Conferencia Latinoamericana de Informática*, page 8p. (Electronically Published), Montevideu, Uruguai.

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

[Macedo et al., 2003] Macedo, A. A., Truong, K., Camacho-Guerrero, J. A., and Pimentel, M. G. C. (2003). Automatically sharing Web experiences through a hyperdocument recommender system. In *ACM Conference on Hypertext and Hypermedia*, pages 48–56.

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

[Obras Raras, 2009] Obras Raras (2009). Obras Raras. Internet (Visited: 27/Jan/2009). `http://www.obrasraras.usp.br`.

[Parapar et al., 2009] Parapar, J., Freire, A., and Barreiro, A. (2009). Revisiting n-gram based models for retrieval in degraded large collections. In *European Conference on IR Research*, pages 680–684, France.

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

[Proin, 2009] Proin (2009). Arquivo Publico do Estado e Universidade de Sao Paulo. Internet (Visited: 27/Jan/2009). `http://www.usp.br/proin`.

[Rios et al., 2010] Rios, I., Britto, Jr., A. S., Koerich, A. L., and Oliveira, L. E. S. (2010). Evaluation of different feature sets in an ocr free method for word spotting in printed documents. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 52–56, New York, NY, USA. ACM.

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

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

[Taghva et al., 1996] Taghva, K., Borsack, J., and Condit, A. (1996). Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing Management*, 32(3):317–327.

[Tan et al., 2002] Tan, C. L., Huang, W., Yu, Z., and Xu, Y. (2002). Imaged document text retrieval without OCR. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):838–844.