Journal of Universal Computer Science, vol. 20, no. 9 (2014), 1154-1173
submitted: 11/2/13, accepted: 28/8/14, appeared: 1/9/14 © J.UCS

Capturing and Relating Multilingual Clinical Cases

Renato de Freitas Bulcão-Neto[1], **José Antonio Camacho Guerrero**
(Innolution Sistemas de Informática
Ribeirão Preto-SP, Brazil
renato@inf.ufg.br, jose.camacho@innolution.com.br)

Paulo Schor, Alessandra Stanquini Lopes
(Departamento de Oftalmologia
Universidade Federal de São Paulo, São Paulo-SP, Brazil
pschor@pobox.com, ale_epm72@yahoo.com.br)

Márcio Branquinho Dutra
(Programa Interunidades de Pós-Graduação em Bioinformática
Universidade de São Paulo, Ribeirão Preto-SP, Brazil
mdutra@gmail.com)

Alessandra Alaniz Macedo
(Departamento de Computação e Matemática
Universidade de São Paulo, Ribeirão Preto-SP, Brazil
ale.alaniz@usp.br)

Abstract: Recent studies reveal that the Internet use has grown tremendously in the past few years, most rapidly in non-English-speaking regions. However, this scenario creates a demand for innovative information retrieval services to better support a world wide community. This paper presents the *MedLink* linking service, which automatically identifies semantic relationships among multilingual clinical cases and makes them available to users as hyperlinks. As a proof of concept, we also present an experiment relating multilingual clinical cases in Ophthalmology, where the relationships created by *MedLink* were qualitatively analyzed by a Faculty with strong Ophthalmology background. Analysis results are described in terms of the completeness and the fidelity of the relationships created, which can be most useful in a globalized world for several purposes including research, teaching, and presurgical decision making.

Key Words: Cross-language Information Retrieval, Semantic Relationships, Multilingual Web, Medical Informatics

Category: H.3, H.4

[1] Mr. Renato Bulcão-Neto worked for Innolution Sistemas de Informática by the time of this research was developed. Nowadays, the current research is still being conducted by the Innolution company. Currently, Mr. Renato Bulcão-Neto is an associate professor in the Institute of Informatics at Federal University of Goiás, Brazil.

1 Introduction

Doctor Mario Amato is member of a multidisciplinary healthcare team which discusses presurgical and postsurgical information of clinical cases during a medical grand round (or clinical meeting). From an updated queue of clinical cases, Dr. Amato chooses three cases to be presented in the next medical grand round. Each clinical case has a particular physician in charge of, who collects reports, slides, imaging exams, videos, etc.

Minutes before a grand round takes place, Dr. Amato and other team members upload all clinical cases material into a particular software, which automatically generates web documents including the material itself and the sequence of users' interactions with that material (e.g. forward and backward slides navigation and handwritten digital ink upon medical images).

During the medical grand round, Dr. Amato often unsuccessfully tries to remember clinical cases similar to the ones being presented, what is explained due to the size of the history of clinical cases presented in the past as well as the large amount of information discussed in medical grand rounds in general.

That collaborative discussion is very important because it leads healthcare team members to share their individual experiences and clinical and scientific evidence-based knowledge towards surgical decision making. Dr. Amato then may open a web browser for seeking clinical cases related to those presented, which could be useful for his own assessments, or even for the collaborative diagnostic conclusion.

However, there is a plenty of clinical cases information on the Web, what may demand from Dr. Amato and his colleagues a stressful process of searching and browsing. Moreover, they may want to formulate queries (i.e. using keywords or a reference document) in their native languages as well as retrieve clinical cases in multiple different languages.

Overall, Dr. Amato's scenario illustrates that there is a demand for software services that identify relationships among documents stored in independent multilingual repositories on the Web as a means of relieving users of reading most documents to find out such relationships. This scenario suggests studies in Information Retrieval (IR) in general, and Cross-Language IR (CLIR) for Medical Informatics in particular. CLIR allows to retrieve information written in a language different from the language of a user query [Peters, 2000].

However, major CLIR techniques exploit translation dictionaries with lexical coverage constraints of the dictionary adopted such as in [McEwan et al., 2002] and [Xu et al., 2002]. Besides, dictionaries are often hand coded towards improving results, what may require much effort to represent large collections.

Considering our IR and CLIR approaches, we developed the *LinkDigger* [Macedo et al., 2008] software framework, which automates the definition of semantic relationships among multilingual textual documents in general, and web documents in particular [Macedo et al., 2005]. The core of *LinkDigger* is the Latent Semantic Indexing (LSI) technique [Furnas et al., 1988], which automatically organizes domain-independent text objects into a semantic structure appropriate for matching. LSI tries to overcome usual problems of lexical approaches such as different words with multiple meanings (polysemy) and identical or at least similar meanings (synonymy).

In this paper, we present a case study in which the *LinkDigger* framework is the core of a multilingual web service called *MedLink*, which is integrated to a medical grand round documenter system called *ArcaMed GRound* [Bulcão-Neto et al., 2008b] as described in Doctor Amato's scenario.

First, we describe how the *MedLink* service collects clinical cases information from *ArcaMed GRound*. Second, we show how that service provides grand rounds' participants with multilingual clinical cases semantically related to the cases being presented. Our data set includes multilingual clinical cases collected from remarkable journals in Ophthalmology on the Web.

Considering Doctor Amato's scenario, we evaluated *MedLink*'s performance in terms of the harmonic mean measure to find the best filtering threshold evenly weighting precision and recall of this service. Besides, the relationships created by *MedLink* were subject of a qualitative evaluation performed by a Faculty with a broad knowledge of Ophthalmology. Regarding completeness and fidelity of such relationships, the quality of results was influenced by some factors, e.g. the size and the heterogeneity of the collection, although it covers a common knowledge area (i.e. Ophthalmology).

This paper is organized as follows. Section 2 reviews CLIR-oriented techniques and related work. Section 3 presents systems software which support the creation of semantic relationships performed by *MedLink*, which is described in Section 4. Section 5 describes and analyzes an experiment with multilingual clinical cases. Section 6 summarizes conclusions and future work.

2 Related work

Although English is the dominant language on the Web, there has recently been a strong demand for multilingual services [Chung, 2008] to support the astonishing increase in the number of non-English-speaking users, who rely on their native languages to seek information on the Web. However, cross-language information retrieval has been a most debated subject for decades.

In the 1970's, Salton's pioneering CLIR approach exploited a thesaurus which took terms for each query term and translated it [Salton, 1969]. The success of this approach over multilingual collections was due to the fact that the thesaurus was created by hand for a specific collection, what avoided ambiguity problems. On the other hand, this approach is limited by the thesaurus vocabulary.

Later, a bilingual CLIR/LSI method was defined with two sets of training documents in which a set of documents is the translation of the other set [Dumais et al., 1996]. Both sets were used to build the initial LSI matrix: lines correspond to words in both languages, and columns represent documents. As a result, subsets of bilingual semantic spaces were defined. New documents could be included in each subset, and independent-language queries could retrieve documents with no translation because all documents were already represented.

As an evolution of Salton's hand coded dictionary, an evolutionary programming approach was developed to optimize query translation combining a machine-readable dictionary with parallel collections [Davis and Dunning, 1995]. Researchers advocate that CLIR based on dictionaries are more effective than using parallel corpus because the former has a wider variability of words in comparison to the latter [Mori et al., 2001]. However, a large scale CLIR system can be built using an enough number of bilingual documents, when no appropriate dictionary is available.

A CLIR technique was developed in which an automatic thesaurus construction method extracted term relationships from the link structure of websites [Chen et al., 2003]. A research goes further by proposing a method for automatically creating and validating candidate Japanese transliterated terms of English words, which are both languages with a few inter-language cognates [Qu et al., 2003].

Researchers advocate that word-by-word translation can not be relied on when a comparable corpus or a parallel corpus is available, and they also argue that comparable corpora contain sets of topically-related documents written in different languages [Lavrenko et al., 2002].

In the context of multilingual medical information retrieval, a dictionary-based approach was proposed in which complex word forms constitute equivalence classes of subwords [Markó et al., 2007]. These semantically minimal units capture intralingual and interlingual synonymy. The disambiguation process relies on a probabilistic model which accounts for co-occurrence information from large textual resources. The evaluation process consists of two different standard test collections for the medical domain in six different languages including English, Portuguese and Spanish.

Researchers [Andrade et al., 2007] developed a method which supports the maintenance of the Markó et al.'s multilingual medical subword repository. Their objective is to automate the errors detection process in the thesaurus content by triggering the lexicographic activities with a ranked list of potential problems generated by the comparison of the semantic extract of comparable corpora.

Regarding inter-clinical case similarity metrics, an abstracting-based approach was developed in which abstracted patient-specific features from medical records were used to improve an information-theoretic measurement [Cao et al., 2008].

The metric proposed, using a combination of abstracted disease, finding, procedure and medication features, achieved good results to experts.

With similar motivation, we have defined comparable corpora using LSI by considering stems of words to automatically compose a vocabulary — instead of using a training set as proposed by Dumais et al. or Markó et al.'s multilingual lexicon and thesaurus. Conversely to Cao et al.'s work, we computed classic information retrieval metrics such as the harmonic mean measure F (or F-measure) [W. M. Shaw et al., 1997].

3 Background

This section describes two systems software used in this work towards capturing and relating clinical cases in different languages: (a) the *ArcaMed GRound* system, which captures medical information shared by healthcare personnel in clinical meetings; and (b) the *LinkDigger* framework, which semantically relates information — captured by the former — available as image documents [Bulcão-Neto et al., 2010] or pure text [Macedo et al., 2002].

3.1 The ArcaMed GRound system

Medical grand rounds are collaborative discussions of clinical cases of patients towards sharing clinical and scientific evidence-based knowledge, individual experiences, and responsibilities. Based on an ethnographic study of medical grand rounds at a university hospital [Bulcão-Neto et al., 2008b], the *ArcaMed GRound* system was developed to address the lack of computational support to automatically document grand rounds discussions.

It captures and registers patients' clinical material and a set of interactions occurred during a medical grand round including navigation, measurements and ink and textual notes upon imaging exams. Every clinical case information is represented in accordance with an XML document structure as in Example 1.

```
<!--                   Example 1                  -->
<clinicalMeeting meetingNumber="">
  <!-- grand round metadata -->
  <clinicalCase>
    <caseDescription> ... </caseDescription>
    <!-- patient metadata -->
    <caseWorkflow>
      <caseEvaluation caseSequence="1">
        <caseMaterial>
          <document docID="" src=""/>
        </caseMaterial>
        <visitSetting>
          <visit visitID="1" docID="" duration=""/>
        <visitSetting>
        <evalDescription> ... </evalDescription>
      </caseEvaluation>
    </caseWorkflow>
    <caseDiagnosis> ... </caseDiagnosis>
  </clinicalCase>
</clinicalMeeting>
```

Lines 3–5 give additional data about a clinical case including a long description and patient's metadata. Lines 6–16 represent the core description of a clinical case, which is usually assessed following a workflow in which every doctor uploads their material (lines 8–10).

During the grand round, every physician presents her documents, and all visit-related information to a document is also registered on the XML excerpt of a clinical case (lines 11–13).

When finishing documents presentation, a physician is able to make a textual assessment related to the current clinical case (line 14) using an embedded text editor. A particular physician finally may register the conclusion of a clinical case (line 17) after all physicians make their own assessments.

All material captured is then automatically transformed and formatted as Web accessible documents for later access [Bulcão-Neto et al., 2007]. These web documents include:

– the material presented (e.g. clinical history, medical reports, image-based examinations, etc.);

– physicians' interactions with that material such as navigation and delimitation of regions of interest upon image-based exams [Bulcão-Neto et al., 2008a];

– physicians' assessments and collaborative diagnostic conclusions of every clinical case discussed during a grand round.

Regarding that documentation, the goal is to enhance and simplify physicians' tasks including the review of particular clinical cases for teaching, research or legal purposes, presurgical decision making, and surgical procedures.

3.2 The LinkDigger framework

In order to find out whether two or more documents are related or not, people usually must read and analyze the content of all documents, what may cause a cognitive overhead to people when the collection of documents is very large. For that reason, the *LinkDigger* framework [Macedo et al., 2008] has been built to automate the identification of relationships among documents by extracting the latent semantics from every document of the collection.

The *LinkDigger* framework architecture is depicted in Figure 1, which describes the orchestration of its two main modules: (a) collect and preprocessing module and (b) linking module.

The *collecting module* gathers information from different sources of textual data[2] including remote locations. As a result, an inverted file of words collected and their respective *tf* frequencies [Baeza-Yates and Ribeiro-Neto, 1999]

[2] Currently, the TXT, DOC, DOCX, RTF, PPT, PPTX, ODT, HTML, XML and PDF formats are supported.

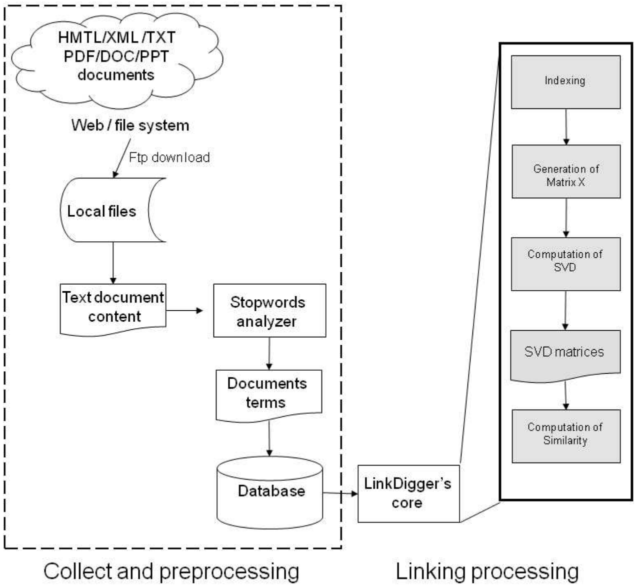

Figure 1: LinkDigger framework architecture based on two modules: (a) collect and preprocessing and (b) linking modules.

are generated. In order to clarify this collect phase in a multilingual context, an example is presented describing a set of selected clinical cases in Ophthalmology gathered from "The Scientific Electronic Library Online" (Scielo) from Brazil and Spain. Ten documents (1 E1, 1 S1 and 8 P1) were randomly selected from those collections to illustrate our example as in Figure 2.

An inverted file is then built, as shown in Figure 3, where lines describe a particular word collected, its frequency in the whole collection, and its frequency in each document (represented by ID) of the collection. The syntax of each line is as follows:

$$word, < word\_freq\_collection,\ (word\_freq\_doc\_1, doc\_1), ..., (word\_freq\_doc\_N,\ doc\_N) >.$$

The *preprocessing module* aims to discard the maximum of non-relevant words from documents. First, the respective language[3] is automatically identified for every word collected. Next, the removal of words that do not semantically represent the documents they belong to: the so-called stopwords (e.g. articles and prepositions). Accentuation and special characters are also removed from each word collected. The frequency of each remaining relevant word is then recalculated for each document of the collection.

[3] Currently, English-, Portuguese- and Spanish-written documents are supported.

Example of Text Data: Titles of some clinical cases in Ophthalmology

E1 - Unilateral Parafoveal Retinal Telangiectasis

P1 – Macrovaso retiniano congênito

P2 – Oftalmomiiase interna posterior relato de dois casos de larva viva no espaço sub retiniano

P3 – Síndrome de Brown bilateral associada com hipermobilidade articular benigna

P4 – Ectrópio palpebral em portador da síndrome de Down e conjuntivite alérgica

P5 – Ectrópio congênito relato de três casos e revisão de literatura

P6 – Córnea verticilata marcador clínico da doença de Fabry

P7 – Esotropia do adulto durante o período gestacional

P8 – Uso de corticóide sistêmico e intravítreo na inflamação secundária a cisticercose intra ocular

S1 – Necrosis retiniana aguda por virus herpes simplex tipo 1 dos años después de meningoencefalitis presuntamente herpética

Figure 2: Titles of 10 clinical cases in Ophthalmology collected from the Brazilian and Spanish Scielo collections, where are found scientific papers written in English, Portuguese and Spanish represented as E1, P1...8 and S1, respectively.

andar	<8,(1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
curriculum	<9,(1,E1), (1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
impressa	<8, (1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
posterior	<26, (1,E1),(3,P1),(15,P2),(4,P4), (1,P8),(2,S1)>
posteriores	<4, (2,P1),(1,P8),(1,S1)>
posteriormente	<1, (1,P1)>
realiza	<2,(1,P4),(1,S1)>
realizada	<12, (1,E1),(1,P1),(2,P2),(2,P5),(3,P6),(3,P8)>
realizadas	<4, (1,P2),(1,P3),(1,P5),(1,P6)>
realizado	<18, (1,P1),(1,P2),(4,P3),(1,P4),(5,P5),(2,P6),(4,P8)>
realizados	<2, (1,P6),(1,P7)>
realizanet	<1,(1,P7)>
realizar	<2,(1,P3),(1,S1)>
realizó	<1,(1,S1)>
realizou	<1,(1,P7)>
references	<1,(1,E1)>
referencia	<1,(1,P6)>
referencias	<10, (1,P1),(1,P2),(1,P3),(1,P4),(2,P5),(1,P6),(1,P7),(1,P8),(1,S1)>
scientific	<9, (1,E1),(1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>

Figure 3: Inverted file with words and their corresponding frequencies from Scielo's clinical cases. For instance, the word “realizadas” appears 4 times (1 in P2, 1 in P3, 1 in P5 and 1 in P6), whereas the words “realizó” and “realizou” appear only once in S1 and P7, respectively.

The next step is the execution of the stemming technique, which reduces each word to its linguistic root (or stem term) based on the language detected. The final step is the calculation of frequency of stem terms as the sum of the *tf* frequencies of all words reduced to a particular stem term. The result is an inverted file of stem terms with their respective frequencies as in Figure 4.

andar	<8,(1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
curriculum	<9,(1,E1), (1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
impress	<8, (1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>
posterior	<31, (1,E1),(6,P1),(15,P2),(4,P4),(2,P8),(3,S1)>
realiz	<43, (1,E1),(2,P1),(4,P2),(6,P3)(2,P4),(8,P5),(7,P6),(3,P7),(7,P8),(3,S1)>
referenc	<12,(1,E1), (1,P1),(1,P2),(1,P3),(1,P4),(2,P5),(2,P6),(1,P7),(1,P8),(1,S1)>
scient	<9, (1,E1),(1,P1),(1,P2),(1,P3),(1,P4),(1,P5),(1,P6),(1,P7),(1,P8)>

Figure 4: Inverted file after preprocessing. The number of terms is reduced since they are grouped such as "references" (English), "referencia" (Portuguese) and "referencias" (Spanish). As identical stem terms are identified (e.g. the stem "referenc" appears 12 times in the whole collection) it approximates stems of languages and reduces the list of terms and the size of matrices handled next.

The *linking processing module* is based on the LSI (Latent Semantic Indexing) technique [Furnas et al., 1988], which distinguishes stem terms that better represent the semantics of a document. Every stem term is given an appropriate weight as its *tf* frequency relative to the frequency of the same term in the whole document collection (*idf* measure). Three term-weighting schema are currently supported by *LinkDigger* [Baeza-Yates and Ribeiro-Neto, 1999, W. M. Shaw et al., 1997, Salton and Buckley, 1988]:

$$w_{ij} = \left(\frac{tf(ij)}{\max tf(ij)}\right) \times \left(\log\left(\frac{N}{n_i}\right)\right) \quad \textbf{(1)}$$

$$w_{ij} = (1 + \log(tf(ij)) \times \log(\frac{N}{ni})) \quad \textbf{(2)}$$

$$w_{ij} = \frac{tf(ij) \times \log(\frac{N}{n_i})}{\sqrt{\sum_{k=1}^{T} (tf(kj))^2 \times (\log(\frac{N}{n_k}))^2}} \quad \textbf{(3)}$$

In all term-weighting schemes, the *idf* factor is calculated as $\log(\frac{N}{n_i})$, where N is the number of documents in the collection, and n_i is the number of documents containing the term k_i. Terms appearing in all documents have a high weight; consequently their *idf* factors are near to zero. Therefore, those formula schema are different in terms of the normalization of tf.

In the scheme **1**, the normalization is defined as $\frac{tf(ij)}{\max tf(ij)}$. Terms that frequently occur in a document have a high weight, and consequently tf is close to 1. In the scheme **2**, the normalization is given in the fraction denominator.

By calculating log of tf of the term and adding 1 to avoid terms have zero weight). Finally, in the scheme **3**, the normalization is accomplished using the log function, i.e. the average of weights.

The next step of the linking processing is the generation of a term-document matrix called matrix X, in which lines represent stem terms and columns represent documents. As shown in Figure 5, entries in the matrix X represent the frequencies of a stem term in each document (e.g. E1, P1, ..., P8, S1).

Terms	E1	P1	P2	P3	P4	P5	P6	P7	P8	S1
andar	0	1	1	1	1	1	1	1	1	0
curriculum	1	1	1	1	1	1	1	1	1	0
impress	0	1	1	1	1	1	1	1	1	0
posterior	1	6	15	0	4	0	0	0	2	3
realiz	1	2	4	6	2	8	7	3	7	3
referenc	1	1	1	1	1	2	2	1	1	1
scient	1	1	1	1	1	1	1	1	1	0

Figure 5: Matrix X with stem terms and respective frequencies.

The matrix X is decomposed into three component matrices T, S and D' using Single Value Decomposition (SVD), which is part of the LSI theory. 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.

The k most important dimensions (the highest values in matrix S) are selected, and all other factors are discarded. We have extended SVD by automatically calculating the value of k given a lost of 30%. The reduced dimensionality solution generates a vector of k real values to represent each document. SVD provides reduced rank-k approximation for the column and row space of a matrix X for any value of k. Finally, the semantic reduced matrix $\hat{X}$ is generated by multiplying the three reduced component matrices, as shown in Figure 6, as a previous step for obtaining the similarity level among documents.

The similarity concept is based on closeness of stem terms into a semantic space built according to the co-occurrence of all stem terms in a document collection instead of simple lexical matching. In order to obtain the similarity level between documents given a particular stem term, the current version implements the cosine [Salton and Lesk, 1968] measure over each pair of document vectors extracted from the semantic matrix $\hat{X}$ (e.g. Figure 6).

Considering that (i) the unitary vectors $vec(i)$ and $vec(j)$ are assumed to be orthonormal, and (ii) the t unitary vectors $vec(i)$ form an orthonormal basis for a

Terms	E1	P1	P2	P3	P4	P5	P6	P7	P8	S1
andar	-0,14	0,19	0,59	0,17	0,47	0,46	-0,33	0,81	0,09	0,10
curriculum	-0,38	-0,40	1,08	0,04	0,93	1,03	-0,35	1,68	-0,05	0,00
impress	-0,07	0,41	0,71	1,61	0,74	2,08	3,66	0,87	4,93	0,95
posterior	0,15	-0,22	0,16	1,69	0,14	0,77	1,63	0,27	0,01	0,80
realiz	-0,19	-0,20	0,54	0,02	0,47	0,51	-0,17	0,84	-0,02	0,00
referenc	0,38	2,66	0,75	4,97	0,20	-0,01	-0,13	-0,03	2,04	2,35
scient	0,12	2,14	1,96	2,21	1,51	2,59	1,77	2,34	1,99	1,37

Figure 6: Matrix $\hat{X}$ for our scenario of multilingual clinical cases.

t-dimensional space, documents and queries are represented as weighted vectors. As a result, documents and queries (or solely documents) can be compared using distance calculation by means of cosine similarity between weighted vectors, as follows:

$$\begin{aligned} Sim(d_j, q) &= cos(\Theta) \\ &= \frac{\overrightarrow{d_j} \bullet \overrightarrow{q}}{|\overrightarrow{d_j}| \times |\overrightarrow{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2{}_{i,q}}} \end{aligned} \tag{1}$$

since $w_{i,j} > 0$ whenever $k_i \in d_j$ and $w_{i,q} >= 0$ whenever $k_i \in q$. $w_{i,q}$ is the weight associated with each term of the query $\overrightarrow{q} = (w_{1,q}, w_{2,q}, ..., w_{t,q})$ and $w_{i,j}$ associated with each term k_i is is associated a unitary vector $vec(i)$. Similarity calculation from (1) also promotes ranking of documents according to cosine results. Figure 7 illustrates document d_j and query or document q being compared using cosine measure ($cos(\Theta)$).

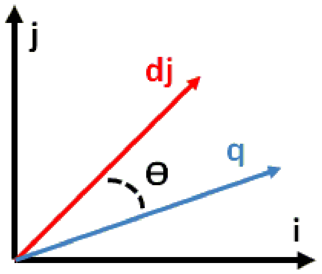

Figure 7: Graphical illustration of cosine measure: cosine values farther to zero mean the highest similarity levels between a pair of documents.

Considering the matrix X in Figure 5, documents E1 and P1, which are written in different languages (English and Portuguese, respectively), have co-occurred terms. A simple query may return both documents with considerable similarity level. Comparatively, after our automatic processing, E1 and P1 are presented as semantic related clinical cases with high similarity (cosine = 96% considering the matrix $\hat{X}$ in Figure 6). This fact was also confirmed by domain specialists, then it would be possible a physician studying the clinical case E1 checking clinical case P1.

The current version allows to query a document collection using keywords or a reference document. Both types of queries are implemented as a column vector of stem term frequencies that can be compared against all columns of matrix X. Considering the matrix X in Figure 5 and a query using the keyword "realiza", all documents (E1, P1...8, S1) would be retrieved with high similarity because it appears in all document collection (i.e. the stem term "realiz"). On the other hand, after defining the semantic matrix $\hat{X}$ in Figure 6, the same query would not retrieve documents with high similarity. Almost all cells for this term are near to zero or negative, what suggests that "realiz" may be also discarded.

In order to filter query results and return the best matches, upper and lower threshold similarity levels are used . If documents are already collected and stored, results are instantly extracted from the list of relationships generated.

4 The MedLink service

The *MedLink* web service provides the *ArcaMed GRound* system with two basic operations: one for collecting clinical cases information, and one for querying relationships among clinical cases (addDocument and getRelationships, respectively). When a clinical case is registered on the *ArcaMed GRound* database, the same contents is also stored on a particular XML document. This way, it is possible to relate clinical cases as textual documents in general.

Example 2 illustrates a clinical case collected from the Brazilian Scielo — note that the most important information of clinical cases is its textual description. For safety reasons, as corroborated by physicians, the respective XML documents of their clinical cases are locally collected and afterwards sent to *MedLink* by invoking the addDocument operation.

```
0 <!--                  Example 2                 -->
1  <clinicalCase>
2    <caseDescription>
       I.F.M., sexo masculino, pardo, 50 anos de idade,
       procurou assistência oftalmológica para consulta de
       rotina referindo dificuldades visuais para leitura.
       Referia ainda diminuição progressiva da acuidade
       visual ao longo da vida e perda da visão central do
       olho direito de maneira progressiva havia 10 anos ...
     </caseDescription>
3  </clinicalCase>
```

Figure 8: MedLink returns to ArcaMed GRound a list with two clinical cases in Spanish (foreground) related to a case in Portuguese (foreground).

During a medical grand round, a physician may request clinical cases related to the case being discussed through the *ArcaMed GRound* user interface, as depicted in Figure 8 (background). This action invokes the getRelationships operation of *MedLink*, which returns an XML document as in Example 3.

Example 3 identifies the current clinical case (line 2) and the corresponding list of related cases (lines 3–5) ordered by similarity level (line 4). That XML document is then transformed to a web page — Figure 8 (foreground) — in order to be presented to *ArcaMed GRound* users.

```
0 <!--                   Example 3                 -->
1  <casesRelated>
2    <document docId="7" src="">
3      <related docId="117" src="">
4        <similarityLevel>0.615107855489</similarityLevel>
5        <title>Distrofia macular anular benigna</title>
6      </related>
7    </document>
8  </casesRelated>
```

Physicians' assessments and the diagnostic conclusion of a particular clinical case are also target data for relating clinical cases. When a grand round finishes, the XML documents of every clinical case presented in the most recent grand round are then re-sent to *MedLink* in order to re-calculate all relationships between clinical cases (including assessments and conclusions).

5 Experimenting on multilingual information

In previous sections, we described a web linking service called *MedLink*, which makes use of information retrieval techniques implemented in our *LinkDigger* framework. This section presents an experiment in which *MedLink* creates hyperlinks among multilingual clinical cases sourced from remarkable online journals in Ophthalmology. The whole process towards creating relationships among clinical cases including the phases of collect, preprocessing, latent semantic indexing, and computation of similarity is described. The same clinical cases were manually interrelated as a result of a qualitative and exhaustive evaluation performed by a researcher with a broad knowledge of Ophthalmology. In order to evaluate *MedLink*´s results, three performance measures broadly accepted in the IR literature were computed and compared to the researcher's results.

5.1 Test data

Test data includes 145 clinical cases in Ophthalmology collected from the Brazilian (`http://www.scielo.br`) and Spanish (`http://scielo.isciii.es`) Scielo. This collection is from remarkable online journals in Ophthalmology in Brazil and Spain, which includes *Arquivos Brasileiros de Oftalmologia*, *Revista Brasileira de Oftalmologia*, and *Archivos de la Sociedad Española de Oftalmología*. The whole collection is composed of 69 clinical cases in Brazilian Portuguese, whereas the remaining (76) are in European Spanish.

Every clinical case (or paper) was downloaded from Scielo and formatted as input data through the *ArcaMed GRound* user interface for the preparation of medical grand rounds. Despite being manually done, this preparation phase is required from physicians to insert clinical cases information even in a real-world medical grand round. From this point, every clinical case information is automatically represented according to the XML file described in Example 1.

5.2 Evaluation

To comprehend analysis results, we describe how *MedLink* was configured to relate multilingual cases:

1. the collect of XML descriptions of clinical cases generated by *ArcaMed GRound*;
2. the execution of preprocessing techniques, e.g. language detection, stop-words, accentuation and special characters elimination, and stemming;
3. and Salton's term-weighting scheme.

From 145 documents (one per clinical case) collected, it were found 23,610 significant words in total. After the stemming process, we found 15,482 stem terms representing a 35% reduction in comparison with the number of significant words, what is normally expected due to the proximity in terms of linguistic roots between Spanish and Portuguese.

Next, *MedLink* performed the term-weighting process to determine the size and the direction of the vectorial representation of each stem term previously found. The matrix X was generated with 15,482 lines (representing stem terms) and 145 columns (representing clinical cases) storing the respective term weight into each document (column). After that, the matrix X was processed using the SVD technique.

We have extended SVD by automatically calculating the value of k given a lost of 30% during the reduction of the matrices. In other words, *MedLink* discards 30% of the highest singular values in the singular matrix S; the remaining are selected to build the semantic reduced matrix $\hat{X}$, whose each pair of document vectors are input parameters for the calculation of cosine-based similarity. Results include a list of pairs of documents and their respective similarity level (in percentage) as depicted in Figure 8 (foreground).

There are two measures extensively used to evaluate the retrieval performance of information retrieval systems — the so-called recall and precision [Baeza-Yates and Ribeiro-Neto, 1999]. The former is the number of the relevant documents which have been retrieved by a system (completeness), whereas the latter is the number of the retrieved documents which are relevant in the whole collection (fidelity). There is often an inverse relationship between recall and precision in which it is possible to increase one at the cost of reducing the other.

In order to evenly weight precision and recall, i.e. good precision with reasonable recall, we calculated the harmonic mean measure F (or F-measure), which assumes a high value only when both recall and precision are also high [W. M. Shaw et al., 1997].

Based on those definitions, we analyzed precision, recall and F-measure obtained from the experiment as depicted in Figure 9. X-axis describes the lower thresholds to filter the number of relationships created, whereas y-axis indicates the values of F-measure, precision and recall of the *MedLink* execution.

We considered relationships created using the whole collection of clinical cases, i.e. Portuguese vs Spanish, Portuguese vs Portuguese, and Spanish vs Spanish. In fact, this represents Doctor Amato's CLIR scenario described in the introduction of this paper. As expected, higher the precision, lower the recall. For instance, a 100% precision corresponds to 0.7% recall — in this case, only four documents are retrieved. In the experiment with clinical cases in Ophthalmology, the best F value is when filtering threshold is 22% — where 138 relationships were created (19% precision and 26% recall).

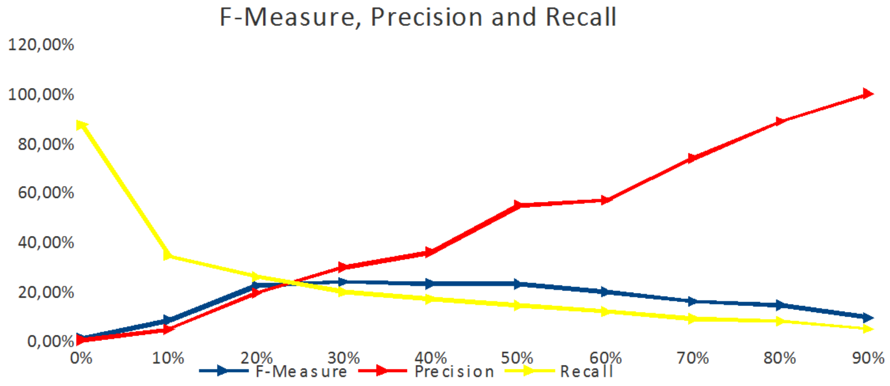

Figure 9: F-measure, precision and recall: relationships between all clinical cases regardless of language.

5.3 Discussion

At a first glance, results may seem not significant, but there are points that worth being discussed. First, the collection size is considerably small, because it is difficult to be eligible to manipulate a medical collection.

Second, the experiment includes only clinical cases from a digital library with free access within a specific period of time where these were published as scientific papers. In order to experiment with unpublished information, it was necessary building a collection only after approval of an Ethics Committee. Moreover, this would be also a complex task to achieve because it demands a permission request for manipulating clinical cases sent to different international institutions since the focus is to relate multilingual information.

After concluded this bureaucratic process, there would still be the need for contacting specialists with multilingual knowledge to evaluate the relationships created by *MedLink*. In the case of this experiment, a Brazilian Faculty with strong knowledge in Ophthalmology, English and Spanish analyzed a small collection with 145 documents which generated the amount of 454 possible relationships. As the collection size increases, more possible relationships need to be analyzed and it is more difficult to compose a multilingual specialist team.

When preparing the collection, it was believed that a collection based on a single knowledge area (Ophthalmology) could be considered a homogeneous repository. From a manual inspection done by the specialist of the experiment, it was found that very different diagnostic impressions from clinical cases indicated groups of knowledge subareas inside the same collection. Experiment results could be probably more expressive if running *MedLink* for each subarea.

Finally, the main focus of this paper was to relate just Spanish and Portuguese clinical cases, but it were found English documents stored in the collection, what surely added noise to experiment results. From 69 clinical cases of the Brazilian Scielo collection, 4 are written in English and include an abstract of the case in Portuguese. Besides, it was found that the remaining clinical cases (141) also include an English-written abstract of the case, what characterizes the document collection used as multilingual.

6 Conclusions and future work

Provided that cross-language information seeking has been recently given special attention, the goal of this paper is to present the *MedLink* service for the automatic creation of relationships between multilingual clinical cases from an automatic documenter system for medical grand rounds.

This approach can carry out experiments from different areas of specialization in Medicine. Pathology and Radiology are good candidates because they usually manipulate a great volume of textual information. However, the highlighted points in Section 5.3 must be taken into account. Shortly, it is strongly necessary both most approved clinical cases from different languages and a group of specialists to create a reference collection.

Dimension reduction may improve recall of information because it reduces the complexity of linear dependence problems and term mismatch. In this work, we apply dimension reduction of LSI when manipulating stem words in matrix X. We do not believe stem words are able to hide latent information. They may avoid term mismatch compared with using complete words.

Regarding the performance measures of the experiment described, the results obtained may be influenced by the homogeneity of the repositories of clinical cases (Ophthalmology), even in different languages. During the experiments, the major problem was the creation of the multilingual medical reference collection, as discussed previously. Our collection had 145 multilingual clinical cases evaluated by specialists. However, it was no possible to find or create a bigger collection due to its multidisciplinarity and multilingualism. It is very difficult to aggregate specialists to attend these two characteristics and the high quantity of relationships.

Finally, as more multilingual information becomes available on the Web, it is expected that mining such external sources for knowledge will play an increasing useful role in relating information problems in general. A coordinator of the Cross-Language Evaluation Forum (CLEF) has been already contacted to investigate the existence of a multilingual medical reference collection.

Recent work has been extending and experimenting the LinkDigger infrastructure towards the automatic generation of relationships among document images using LSI and optical character recognition (OCR) [Bulcão-Neto et al., 2011].

Results have shown the feasibility of LSI relating OCR output even with high degradation.

Future work includes the development of *LinkDigger* components addressing structured documents, fine-grained similarity, collect and transcription of audio files, graphical formalisms to represent relationships among documents, data clusterization at low computational cost[4], and parallelism support of LSI steps.

A formal usability study with physicians should be also performed as future work even though we have followed interfaces of traditional search engines on the Web when developing the *ArcaMed GRound* interface for returning clinical cases related, as in Figure 8 (foreground).

Acknowledgments

We thank the INCT-ADAPTA, CNPq (n. 557976/2008-1, n. 481402/2011-0) and FAPESP (n. 03/07968-9, n. 04/12477-7, n. 05/60729-8, n. 05/60038-5, n. 06/58984-2) Brazilian funding agencies for the invaluable support regarding the ArcaMed and LinkDigger projects. We also specially thank Iuliana, Rodrigo and Juliana, ArcaMed and LinkDigger team members, for their helpful support.

References

[Andrade et al., 2007] Andrade, R. L., Pacheco, E. J., Cancian, P. S., Nohama, P., and Schulz, S. (2007). Corpus-based error detection in a multilingual medical thesaurus. In *Proceedings of the 12th World Congress on Health (Medical) Informatics - Building Sustainable Health Systems*, pages 529–534.

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

[Bulcão-Neto et al., 2010] Bulcão-Neto, R. F., Camacho-Guerrero, J. A., 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 *Proceedings of the 25th ACM Symposium on Applied Computing*, pages 13–17, Sierre, Switzerland. ACM Press.

[Bulcão-Neto et al., 2011] Bulcão-Neto, R. F., Camacho-Guerrero, J. A., Dutra, M., Álvaro Barreiro, Parapar, J., and Macedo, A. A. (2011). The use of latent semantic indexing to mitigate ocr effects of related document images. *Journal of Universal Computer Science*, 17(1):64–80.

[Bulcão-Neto et al., 2008a] Bulcão-Neto, R. F., Camacho-Guerrero, J. A., and Macedo, A. A. (2008a). Automatic documentation of users interactions with DICOM images: A case study in medical grand rounds. In *Proceedings of the VIII Workshop of Medical Informatics*, pages 223–226, Belém-PA, Brazil.

[Bulcão-Neto et al., 2007] Bulcão-Neto, R. F., Macedo, A. A., Wichert-Ana, L., Sankarankutty, A. K., Azevedo-Marques, P. M., and Camacho-Guerrero, J. A. (2007). Prototyping a capture and access application to document medical grand rounds. In *Proceedings of the 13th Braziliam Symposium on Multimedia and Web Systems*, pages 1–7, Gramado-RS, Brazil.

[4] The LSI complexity is $O(mnc)$, 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.

[Bulcão-Neto et al., 2008b] Bulcão-Neto, R. F., Macedo, A. A., Wichert-Ana, L., Sankarankutty, A. K., Azevedo-Marques, P. M., and Camacho-Guerrero, J. A. (2008b). Supporting ethnographic studies of ubiquitous computing in the medical grand round experience. In *Proceedings of the 23rd ACM Symposium on Applied Computing*, pages 1641–1645, Fortaleza-CE, Brazil. ACM Press.

[Cao et al., 2008] Cao, H., Melton, G. B., Markatou, M., and Hripcsak, G. (2008). Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. *Journal of Biomedical Informatics*, 41:882–888.

[Chen et al., 2003] Chen, Z., Liu, S., Wenyin, L., Pu, G., and Ma, W.-Y. (2003). Building a web thesaurus from web link structure. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 48–55, Toronto, Canada. ACM Press.

[Chung, 2008] Chung, W. (2008). Web searching in a multilingual world. *Communications of the ACM*, 51(5):32–40.

[Davis and Dunning, 1995] Davis, M. W. and Dunning, T. (1995). Query translation using evolutionary programming for multi-lingual information retrieval. In *Evolutionary Programming*, pages 175–185.

[Dumais et al., 1996] Dumais, S. T., Landauer, T. K., and Littman, M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proceedings of the Workshop of Cross-linguistic Information Retrieval*, pages 16–23, Zurich, Switzerland. ACM Press.

[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 *Proceedings of the 11th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 465–480, Grenoble, France. ACM Press.

[Lavrenko et al., 2002] Lavrenko, V., Choquette, M., and Croft, W. (2002). Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 175–182, Tampere, Finland. ACM Press.

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

[Macedo et al., 2005] Macedo, A. A., Camacho-Guerrero, J. A., and Pimentel, M. d. G. C. (2005). Bilingual linking service for the web. In *Proceedings of the Symposium on String Processing and Information Retrieval (SPIRE)*, pages 45–48, Buenos Aires, Argentina.

[Macedo et al., 2002] Macedo, A. A., Pimentel, M. G. C., and Camacho-Guerrero, J. A. (2002). An infrastructure for open latent semantic linking. In *Proceedings of the 13th ACM Conference on Hypertext and Hypermedia*, pages 107–116, College Park, Maryland, USA. ACM Press.

[Markó et al., 2007] Markó, K. G., Daumke, P., Schulz, S., Klar, R., and Hahn, U. (2007). Large-scale evaluation of a medical cross-language information retrieval system. In *Proceedings of the 12th World Congress on Health (Medical) Informatics - Building Sustainable Health Systems*, pages 392–396.

[McEwan et al., 2002] McEwan, C. J. A., Ounis, I., and Ruthven, I. (2002). Building bilingual dictionaries from parallel web documents. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 303–323, London, UK. Springer-Verlag.

[Mori et al., 2001] Mori, T., Kokubu, T., and T.Tanaka (2001). Cross-lingual information retrieval based on LSI with multiple word spaces. In *Online Proceedings of the 2nd Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pages 1–8, Tokyo, Japan.

[Peters, 2000] Peters, C., editor (2000). *Workshop of Cross-Language Evaluation Forum*, number 2069 in Lecture Notes in Computer Science, Portugal, Lisbon. Springer.

[Qu et al., 2003] Qu, Y., Grefenstette, G., and Evans, D. A. (2003). Automatic transliteration for japanese-to-english text retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 353–360, Toronto, Canada. ACM Press.

[Salton, 1969] Salton, G. (1969). Automatic processing of foreign language documents. In *Proceedings of the 1969 Conference on Computational Linguistics*, pages 1–28, Morristown, NJ, USA. Association for Computational Linguistics.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513 – 523.

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

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

[Xu et al., 2002] Xu, J., Fraser, A., and Weischedel, R. (2002). Empirical studies in strategies for arabic retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 269–274, Tampere, Finland. ACM Press.