

FlexIR: A Domain-Specific Information Retrieval System

Saïd Radhouani, Claire-Lise Mottaz Jiang, and Gilles Falquet

Abstract—We present a precise search engine adapted to professional environments which are characterized by a domain (e.g. medicine, law, sport, and so on). In our approach, each domain has its own terminology (i.e. a set of terms that denote its concepts: team, player, etc.) and it is organized along dimensions, such as person, location, etc. The dimensions, as described below, are made of concepts and semantic relationships that represent a particular perspective or point of view on the domain. We mainly use the notion of domain dimension to: i) precisely index document content, and ii) develop an interactive interface which allows the user to precisely describe his or her information need and therefore precisely access the document collection.

Index Terms—Information retrieval, domain dimensions, user interface.

I. INTRODUCTION

INFORMATION Retrieval Systems (IRS) are nowadays very popular, mainly due to the popularity of the Web. Most IRS on the Web (also called search engines) are not really domain-oriented: the same techniques are used to index any document. We think that there is a niche for domain-specific IRS: once the document domain is known, certain assumptions can be made and specific knowledge can be used. Users are then allowed to utilize much more precise queries than the usual small set of keywords in use for Web search engines.

In professional environments, IRS should be able to process precise queries, mostly due to its use of a specific terminology, but also because the retrieved information is meant to be part of a user task (diagnose a disease, write a report, etc.). In professional environments, there is also a growing need for accessing information about specific domain documents in many languages and many types of media.

In this paper, we present a precise search engine adapted to professional environments that are characterized by a domain (e.g. Medicine, Law, Sport, and so on). In our approach, each domain has its own terminology (i.e. a set of terms that denote its concepts) and it is organized along dimensions, such as *Person*, *Location*, etc. Dimensions, as described below, are defined by concepts and semantic relationships that represent a particular perspective or point of view on the corresponding domain. We mainly use the notion of domain dimension to:

i) precisely index document content and ii) implement an interactive interface that allows users to precisely describe his or her information need, and therefore precisely access a document collection.

Our main goal through this system is to allow users fluid access to a digital library that contains documents belonging to specific domains, written in different languages, and using different medias. In particular, our system provides the user at all times with a feeling of control and understanding. It therefore provides a keyword search combined with a flexible navigation system. This combination allows a user to select a domain of his interest, build his query, expand and refine it, and select the language and the medias of the search results.

This paper is organized as follows: We first introduce the notion of domain dimensions (Section II). In Section III, we present the main principles of our system interface. Before concluding (Section V), Section IV is dedicated to our system architecture and management.

II. DOMAIN DIMENSIONS

Domain dimensions refer to semantic categories of concepts used to characterize information items (themes) in a specific domain. Each dimension has a name, such as *Team*, *Person*, *Competition*, *Location* in the sport domain; *Pathology*, *Human Anatomy*, *Image Modality*, *Stage of the pathology*, *Type of treatment* in the medical domain, and so on. A dimension is defined by a hierarchy of concepts belonging to the underlying domain. For example, the *Person* dimension may include *Player*, *Referee*, *Coach*, and so on.

The type of the semantic relationship that defines a hierarchy of concepts depends on each domain dimension. For example, the *Person* dimension is defined by the *is-a* relationship (eg. *David Beckham is-a Player*), while the *Human Anatomy* dimension is defined by the *is-part-of* relationship (eg. *Femur is-part-of Leg*).

Our experiments have shown that it is often more convenient and efficient to build *is-a* hierarchies that encompass both the subsumption (generic-specific) and the instantiation relationships. This leads us to consider that every term designates a concept. For instance, *David Beckham* will be considered as a concept and not as an instance (object).

A. Domain Dimensions & Information Retrieval

We use domain dimensions for solving domain-specific precise queries that are characterized by a specialized terminology and a complex semantic structure. In this case, domain

Manuscript received February 3, 2009. Manuscript accepted for publication March 20, 2009.

Saïd Radhouani is with the Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Claire-Lise Mottaz Jiang and Gilles Falquet are with Centre Universitaire d'Informatique, University of Geneva, 7, route de Drize, CH 1227 Carouge, Switzerland



Fig. 1. Dimensions-based search interface of FlexIR

This kind of interface allows flexible ways to access the contents of the underlying collection. For example, from the Human Anatomy dimension, a user can choose to select the Skeleton subcategory, and from this select in turn the Leg Bones subcategory. The user can choose any other dimension, perhaps Pathology and Modality, and from this select the Fracture category, and then group the resulting images by X-ray, MRI, or any other dimension (stage of the pathology, treatment, and so on).

A recent usability study on facet-based interface demonstrated that this kind of interface is very flexible and intermediate in complexity [14]. During this study, a strong majority of participants preferred being allowed to navigate in multiple dimension hierarchies simultaneously rather than one dimension; they felt they were in control and did not feel lost. The approach reduces mental work by promoting recognition over recall and suggesting logical but perhaps unexpected alternatives at every turn.

While dimensions-based search interfaces are used primarily in domain-specific collections, there are many movements to promote larger scale use of metadata more generally (eg. careerone.com.au, eBay, etc.).

We have been investigating how to build an intuitive interface for our dimensions-based IRS. The resulting interface has been developed according to the usability results of and the recommendations presented by Hearst and her group [14].

The interface includes a personalization feature. When selecting a domain, the domain's dimensions are automatically shown, and the user can add other dimensions that are better suited to represent his information need. He can also remove dimensions to avoid cluttering the interface with irrelevant information. Figure 2 schematically shows the basic user interactions (actions) that are used to build up a query.

IV. SYSTEM ARCHITECTURE AND MANAGEMENT

The architecture of our system is presented in Figure 3. There are mainly three steps in our system design: *i*) defining

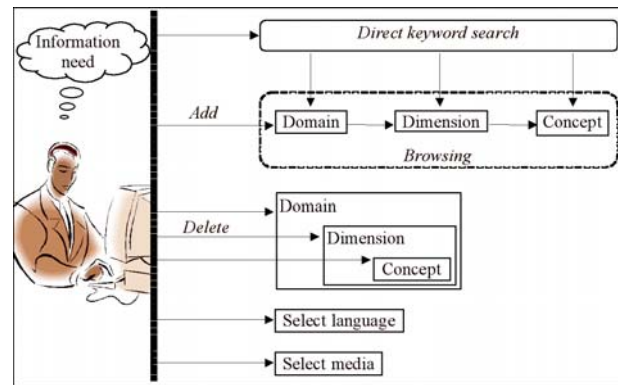


Fig. 2. Basic user actions

the domain dimensions relevant to the given document collection; *ii*) multidimensional document indexing, which matches associate documents to the corresponding domain dimensions; *iii*) external resources preparation. These steps are described in the following sections. We call the resulting system **FlexIR: Flexible Information Retrieval** system.

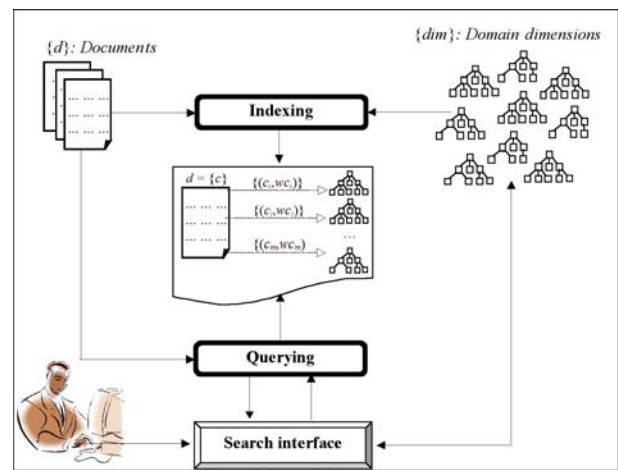


Fig. 3. The architecture of FlexIR system

A. External Resources-Based Domain Dimension Definition

Our aim is to access a multimedia multilingual document collection through a dimensions-based search interface. Instead of using a specific indexing technique for each language and each media, we propose to use a unique technique for all documents independent of their languages or their media type. This technique is based on conceptual indexing that consists in representing documents (queries) by concepts instead of ambiguous descriptors (words extracted from text, features extracted from audiovisual information such as colour, shape, texture, motion, audio frequency, etc). For example, in the UMLS¹ Meta-Thesaurus, the concept denoted in English

¹<http://www.nlm.nih.gov/research/umls/> [visited on March 2008]

ER containing a pivot core (each concept is identified by a unique identifier and denoted by a specific term in each language). Our concept alignment algorithm [19][20] is based on the similarity of concept descriptions (both structural and linguistic) and the similarity of the documents associated to the concepts (definition, description, examples, etc.)

After choosing or building an appropriate ER for a given domain, we define it through the corresponding dimensions. This step consists in selecting, from the ER, the hierarchy of concepts defining each dimension. In a practical perspective, our experiments have shown that it is relatively easy to manually extract a dimension from a vast knowledge resource such as UMLS. We are now addressing the problem of automatic dimensions construction. We have proposed an algorithm for this purpose and the evaluation results are not yet conclusive. Recently, some tentative experiments have been carried out in this direction and the result are promising [18].

V. CONCLUSION

We presented an information retrieval system adapted to professional environments that are characterized by a domain. This system allows user to access in a fluid manner digital libraries that contain documents belonging to specific domains, in different languages, and in different medias. The underlying information retrieval approach is based on the use of dimensions, which refer to semantic categories of concepts used to characterize information items (themes) in a specific domain. Dimensions are used to precisely index documents content and implement an interactive interface that allows a user to precisely describe his or her information need, and therefore precisely access a document collection. We use multilingual external resources to define dimensions and index documents in different languages using concepts instead of terms. Thus, through our interface, a user can formulate his query in his favorite language and access a collection containing documents in several languages. Based on domain dimensions defined through multilingual external resources, our system gives the user at all times a feeling of control and understanding. It therefore provides a keyword search combined with a flexible navigation system, where a user can select the domain of his interest, build his query, expand and refine it, and select the language and the media of his search results.

ACKNOWLEDGMENT

This work was supported by the Swiss National Science Foundation.

REFERENCES

- [1] Hearst, M.A., "Clustering versus faceted categories for information exploration," *Commun. ACM* 49, pp 59-61, 2006.
- [2] Pollitt, A.S., "The key role of classification and indexing in view-based searching," in *Proceedings of the 63rd International Federation of Library Associations and Institutions General Conference (IFLA'97)*, 1997.

- [3] Yee, K.P., Swearingen, K., Li, K., Hearst, M., "Faceted metadata for image search and browsing," in *CHI '03: Proceedings of the conference on Human factors in computing systems*, ACM Press, pp. 401-408, 2003.
- [4] Mäkelä, E., Hyvönen, E., Saarela, S., "Ontogator - a semantic view-based search engine service for web applications," in *International Semantic Web Conference*, pp. 847-860, 2006.
- [5] Mäkelä, E., Hyvönen, E., Sidoroff, T., "View-based user interfaces for information retrieval on the semantic web," in *ISWC-2005 Workshop End User Semantic Web Interaction*, November 2005.
- [6] Sacco, G.M., "Research results in dynamic taxonomy and faceted search systems," in *DEXA Workshops*, IEEE Computer Society, pp. 201-206, 2007.
- [7] Diederich, J., Thaden, U., Balke, W., "The semantic growbag demonstrator for automatically organizing topic facets," in *ACM SIGIR Workshop on Faceted Search*, Seattle, USA, 2006.
- [8] Diederich, J., Balke, W.T., Thaden, U., "Demonstrating the semantic growbag: automatically creating topic facets for faceted dblp," in *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, New York, NY, USA, ACM, pp. 505-505, 2007.
- [9] Radhouani, S. and Falquet, G., "Using External Knowledge to Solve Multi-Dimensional Queries," in *Proc. 13th Intl Conf. on Concurrent Engineering Research and Applications (CE 2006)*, Antibes, IOS Press, Sept. 2006.
- [10] Guyot, J., Radhouani, S. and Falquet, G., "Conceptual Indexing for Multilingual Information Retrieval" in *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 4022, Springer, 2005.
- [11] Radhouani, S., Maisonnasse, L., Lim, J.-H., Le, T.-H.-D. and Chevallet J.-P., "Une Indexation Conceptuelle pour un Filtrage par Dimensions, Expérimentation sur la base médicale ImageCLEFmed avec le métathésaurus UMLS," in *Proc. Conférence en Recherche Information et Applications CORIA'2006*, Lyon, France, 15-17 mars, 2006.
- [12] Radhouani, S., Lim, J.-H., Chevallet, J.-P. and Falquet, G., "Combining Textual and Visual Ontologies to Solve Medical Multimodal Queries," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME 2006)*, Toronto, Canada, July 9-12, 2006.
- [13] Yee, K.P., Swearingen, K., Li, K., and Hearst, M., "Faceted metadata for image search and browsing," in *Proceedings of CHI 2003*, Fort Lauderdale, FL, Apr. 2003.
- [14] Marti Hearst, "Design Recommendations for Hierarchical Faceted Search Interfaces," in *ACM SIGIR Workshop on Faceted Search*, August, 2006.
- [15] Kaki, M., "Findex Search result categories help users when document rankings fail," in *Proceedings of ACM SIGCHI*, Portland, OR, Apr. 2005.
- [16] Radhouani, S., Charhad, M. and Falquet, G., *Multi Point of View Document Categorization: Application to Broadcast News Documents*, Technical report, CUI, University of Geneva, April, 2005.
- [17] Guyot, J., Falquet, G., Benzineb, K., *Construire un moteur d'indexation*, Technique et science informatique (TSI), Hermes, Paris, 2006.
- [18] Stoica E., Hearst M., and Richardson M., "Automating Creation of Hierarchical Faceted Metadata Structures," in *Proceedings of NAACL-HLT*, Rochester NY, April 2007.
- [19] Falquet, G., Mottaz Jiang, C.-L., Ziswiler, J.-C., "Ontology Based Interfaces to Access a Library of Virtual Hyperbooks," in Rachel Heery, Liz Lyon (Eds.) *Research and Advanced Technology for Digital Libraries. Proceeding of the 8th European Conference on Digital Libraries (ECDL 2004)*, Bath, UK, September 12-17, 2004, Lecture Notes in Computer Sciences (LNCS), vol. 3232, Springer, Berlin, Germany, 2004.
- [20] Falquet, G., Nerima, L., Ziswiler, J.-C., "Augmented Hyperbooks through Conceptual Integration," in *Proceedings of the 16th ACM Conference on Hypertext and Hypermedia (Hypertext'05)*, Salzburg, Austria, September 6-9, 2005.