Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Flexible Information Retrieval System for A Dimension Based Search Interface

Raja Sekhar raju*
M .Tech 2nd year, Dept of CSE,
ASCET, GUDUR, India.
raj.merits@gmail.com

Dora Babu Sudarsa
Associate Professor, Dept of CSE
ASCET, GUDUR, India.
dorababu.sudarsa@gmail.com

Abstract— *Search engines are the most commonly used type of tool for finding relevant information on the Internet. However, today's search engines are far from perfect. Typical search queries are short, often one or two words, and can be ambiguous therefore returning inappropriate results. 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. These evaluation initiatives have led to a considerable increase in system performance. Data for evaluation efforts include multilingual corpora, structured data, scientific documents, Web pages as well as multimedia objects. This paper gives an overview of the current activities of the major evaluation initiatives. Special attention is given to the current tracks and developments within TREC, CLEF and NTCIR. The evaluation tasks and issues, as well as some results, will be presented*

Keywords— *Information retrieval, domain dimensions, user interface, flexIR, CLEF TREC, GEOCLEF*

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 W ~b 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 *t}* 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.

II. RELATED WORK

2.1 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 dimensions are used to extract the specialized vocabulary and therefore highlight the relevant elements that contribute to the description of a document (or query) semantic content. For example, through our dimension-based model, a journalist wishing to write a newspaper article can formulate his query as follows:
"Give me documents dealing with the French General who created the security zone during the Balkans conflict". Our

system is able to recognize domain dimensions and use them to precisely answer this query: *Person* (French General), *Location* (Balkans, security zone), *Event* (Balkans conflict).

A relevant document may, for instance, contain the name "Philippe Morillon" without necessarily containing the terms "French" and "General." Thus, from this query, our system can interpret that the journalist is looking for a *Person* who is-a *General originally from* "France" and for a *Location* (Security Zone) that *is-a-part-of* "Balkans." This document cannot be found by a system based on term matching. We therefore use domain dimensions and semantic relationships to precisely interpret users' information needs. We have concluded through a series of experimental evaluations that the use of domain dimensions significantly improves the retrieval performance and outperforms existing approaches that do not take into account domain dimensions [9]. The obtained results encouraged us to implement a search interactive interface where a user can take advantage of domain dimensions during his query process. We therefore address the question of how to integrate domain dimensions during the query process and provide access to users through an interactive interface. In addition to our results, our current research is also motivated by a series of usability studies that find that dimensions (facet)-based interfaces are overwhelmingly preferred over the standard keyword-and-results listing interfaces used in Web search engines [7]. Moreover, a study has shown that information seekers often express a desire for a user interface that organizes search results into meaningful groups in order to help make sense of the results, and to help decide what to do next. A longitudinal study in which participants were provided with the ability to group search results found they changed their search habits in response to having the grouping mechanism available.

2.2. SEARCH INTERFACE

The main idea behind a search interface based on domain dimensions is quite simple. Rather than creating one large category hierarchy, we build a set of category hierarchies, each of which corresponds to a different dimension (facet) relevant to the domain described in the collection to be navigated. This representation is also known as hierarchical faceted categories. In dimension-based search interfaces, each domain has a set of dimensions and each dimension has a hierarchy of concepts. After the dimensions' hierarchies are designed, each document in the collection can be assigned to many concepts from the hierarchies. For example, in the medical domain, the dimension hierarchies can include *Human Anatomy* (Head, Brain, Femur, etc. with *part-of* relationships), *Pathology* (Cancer, fracture, lesion, etc., with is-a relationships), *Image Modality* (MRI, x-ray, ultrasound, etc., with is-a relationships) and so on. Thus, an *MRI* image describing a fracture of a femur might be assigned to (indexed by):

Anatomy > Musculoskeletal System > Skeleton> Bone and
Bones> Bones of Lower Extremity> Foot Bones> Leg Bones> Femur
Pathology > Disorders of Environmental Origin> Wounds
and Injuries > Fractures, Bone > Femoral Fractures
Modality > Diagnostic Techniques and Procedures> Diagnostic
Imaging > Magnetic Resonance Imaging

When a concept within a dimension hierarchy is selected within the interface, all documents that have been assigned to that concept are retrieved (and displayed). When concepts from different hierarchies are selected, the system builds a query that is a conjunct of disjuncts over the selected concepts and their sub concepts.

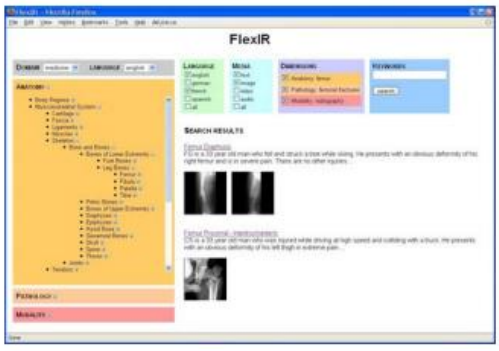

Fig. 1. Dimensions-based search interface of FlexIR

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 Xray, 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 developped according to the usability results of and there commendations 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.

3.SYSTEM ARCHITECHTURE AND MANAGEMENT

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

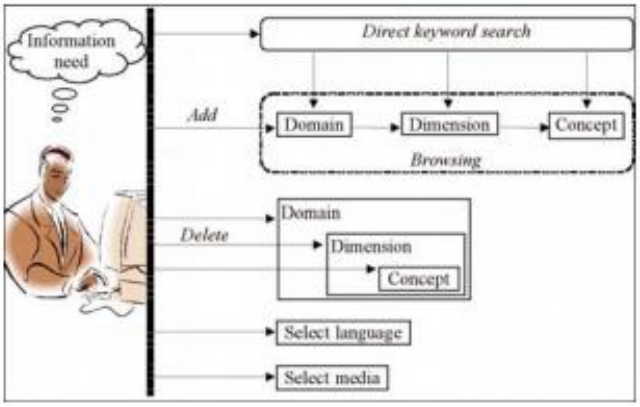

Fig. 2. Basic user actions

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.

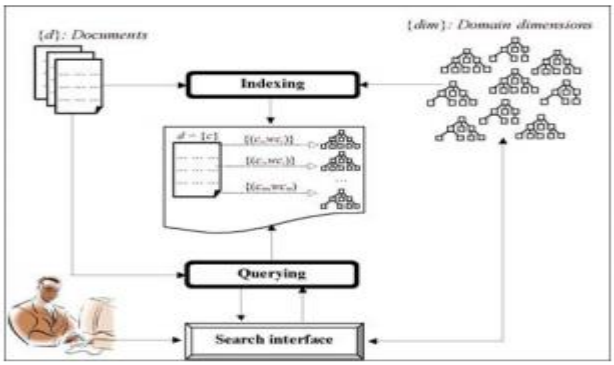

Fig. 3. The architecture of FlexIR system

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).

3.1Multi-Dimensional Document Indexing

In this section the second step is presented. It consists in assigning of each document to its appropriate dimension and consequently to the domains to which it belongs. Let {dom}, {dim}, and {doc} be, respectively, the sets of domains , dimensions and documents present in the collection. Through a conceptual indexing process, each document doc is represented by a set of concepts: doc = {c}. Our approaches for conceptual indexing and the underlying results are detaile din our previous works: multilingual text retrieval [10][11],Image retrieval [12] and video retrieval [16].In order to give the user a list of documents ranked in their order of relevance with respect to his information need, we use a weight schema for weighting all document concepts. Thus, after extracting all concepts from a document doc, each concept c will be given a weight wc that represents its importance in describing the content of doc. The importance of a concept depends on its frequency in the document and on its context (its relationships with the other concepts of the same document). Our weighting method is based on a multidimensional document indexing approach that we defined and evaluated in a previous work [9]. Each document is finally represented by a set of weighted concepts: doc = {(c,

wc)}.The next step is to associate each document with all corresponding dimension hierarchies and the underlying domains.The association between a document doc and a dimension dim is materialised by a link between doc and each concept c ∈ doc∩Cdim). Each link between dom and a concept c ∈ Cdimis labeled by the weight wc (See figure 3). For instance, an Xray image describing a fracture of a femur will be associated respectively with the dimensions Modality, Pathology, and Anatomy. Finally, each document doc implicitely belongs to all domains containing the dimensions to which doc has beenassociated.

3.2External Resources Preparation

The main goal is to select for each domain one or several external resources (ER). The main criteria taken into account when we choose an ER are: that it covers the vocabulary of the domain in different languages, contains lexical structure (association between terms and concepts), and has a hierarchical structure. In some cases, one ER can satisfy all these criteria. For instance, the meta-thesaurus UMLS is an appropriate ER for the medical domain. Nevertherless, when there is no singleer that satisfies all these criterion, we select several ERs so that their fusion gives an appropriate ER for the corresponding domain. For example, we can choose for the International Politic domain independent ERs describing respectively thedimensions Persons, International Events, Locations, and soon.

4. 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 on tainting 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 cans elect the domain of his interest, build his query, expand and refine it, and select the language and the media of his search results..

REFERENCES

[1] Hearst, M.A., “Clustering versus faceted categories for informationexploration,” Commun. ACM 49, pp 59-61, 2006.

[2] Pollitt, A.S., “The key role of classification and indexing in viewbased searching,” in Proceedings of the 63rd International Federationof Library Associations and Institutions General Conference (IFLA’97),1997.

[3] Yee, K.P., Swearingen, K., Li, K., Hearst, M., “Faceted metadata forimage search and browsing,” in CHI ’03: Proceedings of the conferenceon Human factors in computing systems, ACM Press, pp. 401-408, 2003.

[4] M¨akel¨a, E., Hyv¨onen, E., Saarela, S., “Ontogator - a semantic viewbased search engine service for web applications,” in InternationalSemantic Web Conference, pp. 847-860, 2006.

[5] M¨akel¨a, E., Hyv¨onen, E., Sidoroff, T., “View-based user interfaces forinformation retrieval on the semantic web,” in ISWC-2005 WorkshopEnd User Semantic Web Interaction, November 2005.

[6] Sacco, G.M., “Research results in dynamic taxonomy and faceted searchsystems,” 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 Workshopon Faceted Search, Seattle, USA, 2006.

[8] Diederich, J., Balke, W.T., Thaden, U., “Demonstrating the semanticgrowbag: 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 SolveMulti-Dimensional Queries,” in Proc. 13th Intl Conf. on ConcurrentEngineering 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 InformationRepositories: 6th Workshop of the Cross-Language Evaluation Forum,CLEF 2005, Revised Selected Papers, Lecture Notes in ComputerScience, 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, Exp´erimentation sur la base m´edicale ImageCLEFmed avec lem´etath´esaurus UMLS,” in Proc. Conf´erence en Recherche Informationet Applications CORIA’2006, Lyon, France, 15-17 mar

12] Radhouani, S., Lim, J.-H., Chevallet, J.-P. and Falquet, G., “CombiningTextual and Visual Ontologies to Solve Medical Multimodal Queries,” inProc. 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 metadatafor image search and browsing,” in ‘Proceedings of CHI 2003, FortLauderdale, FL, Apr. 2003.[14] Marti Hearst, ”Design Recommendations for Hierarchical FacetedSearch Interfaces,” in ACM SIGIR Workshop on Faceted Search, August,2006.

[14] Kaki, M., "Findex Search result categories help users when documentrankings fail," in Proceedings of ACM SIGCHI, Portland, OR, Apr. 2005.

[15] Radhouani, S., Charhad, M. and Falquet, G., Multi Point of ViewDocument Categorization: Application to Broadcast News Documents,Technical report, CUI, University of Geneva, April, 2005.

[16] Guyot, J., Falquet, G., Benzineb, K., Construire un moteur d'indexation,Technique et science informatique (TSI), Hermes, Paris, 2006.

[17] Stoica E., Hearst M., and Richardson M., "Automating Creation ofHierarchical Faceted Metadata Structures," in Proceedings of NAACLHLT, Rochester NY, April 2007.

[18] Falquet, G., Mottaz Jiang, C.-L., Ziswiler, J.-C., "Ontology Based Interfaces to Access a Library of Virtual Hyperbooks," in Rachel Heery, LizLyon (Eds.) Research and Advanced Technology for Digital Libraries. Proceeding of the 8th European Conference on Digital Libraries (ECDL2004), Bath, UK, September 12-17, 2004, Lecture Notes in ComputerSciences (LNCS), vol. 3232, Springer, Berlin, Germany, 2004.

[19] Falquet, G., Nerima, L., Ziswiler, J.-C., "Augmented Hyperbooksthrough Conceptual Integration," in Proceedings of the 16th ACMConference on Hypertext and Hypermedia (Hypertext'05), Salzburg,Austria, September 6-9,

