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 differnt 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

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 [13]. 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 [15].

B. Related Works

Our goal is to create a domain-specific IRS that takes into account domain dimensions during retrieval process. The research devoted to the concept of dimensions is mainly related to the development of tools for navigating through document collections. These tools are based on the paradigm of research known as faceted search [1][2][3] or view-based search [4][5].

The idea behind the paradigm of faceted research is that, on order to be classified, a document has different characteristics (aspects), each of which can be described by a hierarchy of different concepts [6]. In this way, search results (documents) can be arranged through intuitively (usually) orthogonal facets. For example, in a digital library, results can be grouped by author, year of publication, theme, etc.

The approaches based on faceted search are promising, but their application is limited to a small scale because the facets' construction and the entire process of document annotation are manual [7][8]. Note also that the facets in this case correspond to metadata. Therefore a user is not able to access a document collection by its content. In addition, facets are represented by hierarchies of terms that do not allow access to a collection containing documents in different languages.

In our domain-specific IRS, dimensions are defined by domain concepts that are denoted by terms in different languages. Thus, users can access a document collection by its content and in different languages. Our document-dimension association approach is completely automatic and relies on a conceptual indexing technique [9]. Our dimensions can also be constructed automatically from different existing resources. In fact, we do not manually build complete hierarchies of concepts to define dimensions. Instead, we benefit from (small) existing structures independent of their nature (thesauris, ontologies, etc.) to align them and automatically build dimensions.

III. Dimensions-Based 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 subconcepts.

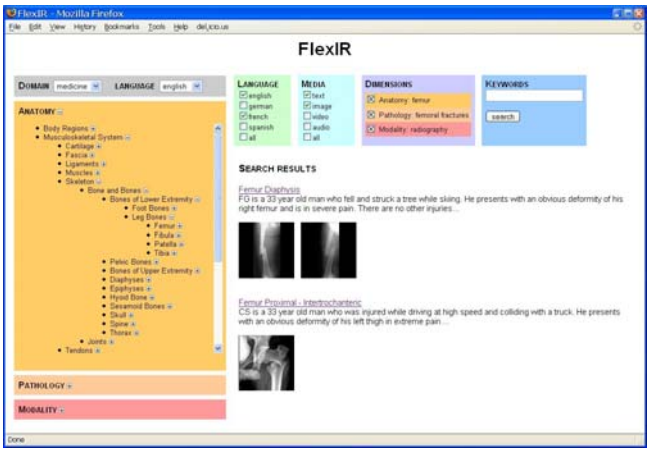

Fig. 1. Dimensions-based search interface of FlexIR

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 developped 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 Architechture and Management

The architecture of our system is presented in Figure 3. There are mainly three steps in our system design: *i*) defining 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**: **Flex**ible **I**nformation **R**etrieval system.

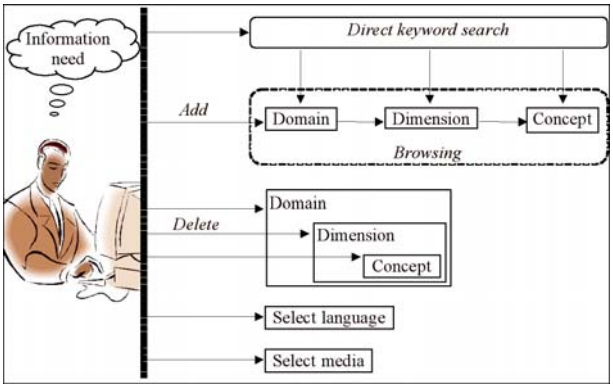

Fig. 2. Basic user actions

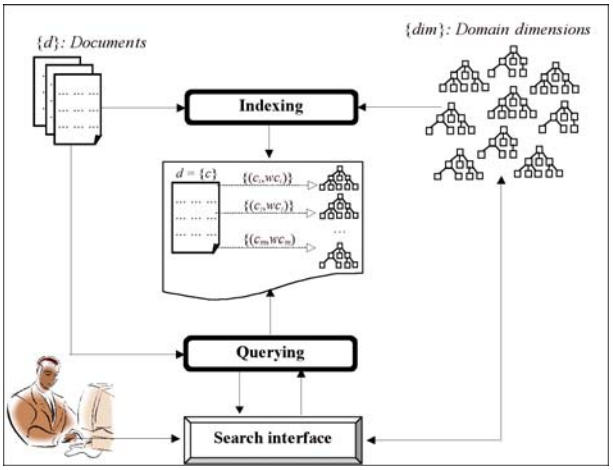

Fig. 3. The architecture of FlexIR system

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 UMLS[1] Meta-Thesaurus, the concept denoted in English

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

by the term "Anterior Cruciate Ligament" is identified by "C0630058". This concept is denoted by a specific term in each language: Fr ⇒ "Ligament Croisé Antérieur", It ⇒ "Legamenti Crociati Anteriori", De ⇒ "Vorderes Kreuzband", and so on. Thus, during the conceptual indexing, all documents dealing with this concept will be represented by the identifier "C0630058". During querying process, a user can express his information need through a textual query formulated in his favorite language. The user's query terms will then be substitued by the corresponding concepts, so that the system retrieves all documents represented by these concepts. The user will be seamlessly given a set of multimedia multilingual documents relevant to his information need.

To establish a viable conceptual indexing platform that can handle domain dimensions, we need a multilingual external[2] resource that must at least have a lexical structure (association between terms and concepts), and a hierarchical structure (relationships between concepts, eg. *is-a* or *part-of*). Ontologies, thesauris, or taxonomies often have these characteristics.

The formal model of an external resource S is a 4-tuple $[C, \leq_C, T, F]$ where:

- C is a set of concepts $\{c_i, \ldots, c_s\}$;
- $\leq_C$ is a partial order on C, called the concept hierarchy;
- $T = \{T_L\}$ is the lexicon of the external resource. It consists of a set of terms $\{t_i, \ldots, t_r\}$ in each language L;
- $F_L \subseteq T \times 2^C$ is a function that associates each term from the language L to the set of concepts it designates. For instance, F_{En}(Anterior Cruciate Ligament) = C0630058 = F_{De}(Vorderes Kreuzband) and F_{De}^{-1}(C0630058) = {Vorderes Kreuzband}.

Once the external resource is chosen, we define the set of dimensions that are relevant to the document collection content. Formally, a dimension dim_i is defined by a hierarchy of concepts as follows: $dim = (root_{dim}, C_{dim})$, where:

- $root_{dim} \in C$ is the root of the hierarchy of concepts defining dim. $F(root_{dim})$ is the name of dim;
- $C_{dim} = \{c \in C \mid c \leq root_{dim}\}$

A dimension dim can be defined either by a subhierarchy of an external resource, or by an entire external resource. For instance, the *Pathology* dimension of the medical domain can be defined by a subhierarchy of the UMLS meta-thesaurus, or by the entire *Disease Ontology*[3].

We finally obtain a set of domain dimensions: $Dim = \{dim_1 \ldots dim_d\}$. Each domain dom belonging to the collection will therefore be defined by the set of dimensions it contains.

B. Multi-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.

[2]"external" because it models knowledge which are not present in the collection to be processed, at least in an explicit and complete form.

[3]http://diseaseontology.sourceforge.net/ [visited on March 2008]

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 detailed in 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 w_c 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 multi-dimensional document indexing approach that we defined and evaluated in a previous work [9]. Each document is finally represented by a set of weighted concepts: $doc = \{(c, w_c)\}$.

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 $c \in doc \cap C_{dim})$. Each link between dom and a concept $c \in C_{dim}$ is labeled by the weight w_c (See figure 3). For instance, an X-ray 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 been associated.

C. External Resources Preparation

The main goal is to select for each domain one or several external resources (ER). The main criteria taken into acount when we choose an ER are: that it covers the vocabularly 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 single ER 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 the dimensions *Persons*, *International Events*, *Locations*, and so on.

The approach we have adopted takes advantage of all possible existing ER independent of their nature and lexical language. In fact, we have developed an algorithm that allows us to align several different ER dealing with the same domain and provide one multilingual ER. The idea is to have first the core of the ER we want to build: a set of concepts denoted by terms and structured in a hierarchy. Then, our algorithm aligns any new ER with the fixed core. The result is a multilingual

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.