What is a « good » Hypertext System for accessing Scientific Literature?

Hermine NJIKE FOTZO

Laboratoire d'Informatique de Paris 6 – LIP6
8 rue du Capitaine Scott, 75015 Paris France
Hermine.Njike-Fotzo@lip6.fr

Abstract

The development and the availability of scientific work in electronic form have changed the way researchers reach scientific literature. To allow retrieval in this literature, there is a need to structure and organize these corpora in a way that reflects some semantic relations between documents. In this paper, we define what would be a good hypertext system for scientific corpora and the useful types of links for the system. We also present some automatic methods for generating scientific hypertext. First results are encouraging and foresee a fully-automatic construction of such systems.

Introduction

With the development of Web, scientific papers are more and more available. Libraries are forsaken with the profit of Web which becomes the main source of access to scientific information. Many of scientific documents collections are loosely structured. Others have been manually structured, most often into hierarchies like those of internet portals (Yahoo, LookSmart, Cora, etc.) or of large collections like MEDLINE: documents are gathered into topics, which are themselves organized into a hierarchy going from the most general to the most specific. There is a great need for structuring and organizing these corpora in a way that reflects semantic relations between documents, so as to offer an intelligent tool to access this information. For now, these relations are indicated mainly via hyperlinks or by organizing documents into concept hierarchies, both being manually developed. It would be necessary to make it possible for the researchers to have first a good summary of the contents of the collection, then to find quickly relevant information and to put the finger on communities in emergence and on the current problems. This requires:

- To structure the overall corpora in order to obtain a representation easily to handle by computers
- To establish links reflecting semantic between the objects of these corpora - these links are currently being created manually
- To make these links dynamic, adapting to the evolution of the corpus by the arrival of new documents or according to the users' navigation.

These needs also appear in the constitution of specialized search engines, the navigation of corpora, for building and maintaining products leading in the field of cultural multi-media, and in a more general way to maintain, enrich and make evolve dynamic corpora on specific topics. Generally, the problem is to move from "flat" corpora to structured corpora, which highlight various types of relation between documents, and which constitute a true base of knowledge.

We are interested in this article in the definition of a good hypertext system to access scientific literature within the framework of active research. We propose to structure this hypertext system by the automatic generation of typed links between elements of the concerned corpus and by the generation of the hierarchies of concepts present in the corpus. Structuring corpora in concepts hierarchies can be view as a method of organization, of summary, of information access but also as the links as a tool for navigation for the large corpora.

The paper is organized as follows. In section 2 we introduce previous related work on automatic collections structuring. In section 3, we will present and justify the useful types of links for scientific corpora. In section 4 we will present some models for generating the selected typed links and give some results. Finally we will end on perspectives generated by this work.

Previous Work

In this section, we present a state of the art concerning the automatic structuring of documents collections. A structural element can be view as any additional dimension brought to the text seen like a simple sequence of words. The principal emerging elements of external structure in the literature of the information retrieval community (IR) are: hyperlinks between documents or parts of documents and the classification of the documents within a hierarchy of concepts going from the most general to the most specific. Structuring a collection is obtained from automatic generation of links between documents or by classifying the documents of this collection within a hierarchy of preset concepts.

Generation of Topics Hierarchies

The generation of hierarchies is a classical problem in information retrieval. In most cases the hierarchies are manually built and only the classification of documents into the hierarchy is automatic.

Clustering techniques have been used to create hierarchies automatically like in the Scatter/Gather algorithm [Cutting et al., 1992], such hierarchies have been used to help navigation or retrieval. Alternatively, hierarchical clustering techniques have been used in many instances for organizing document corpus. All these methods cluster documents according to their similarity. They cannot be used to produce topic hierarchies.

Recently, topic hierarchies more similar to those found in e.g. Yahoo have been proposed. As in Yahoo, each topic is identified by a single term. These term hierarchies are built from "specialization/generalization" relations between the terms, automatically discovered from the corpus. They can eventually be used to create document hierarchies: one document is attached to a term node if this term is characteristic of the document. [Sanderson and Croft, 1999] propose to build term hierarchies based on the notion of subsumption between terms. A subsumption hierarchy reflects the topics covered within the documents, a parent term is more general than its child, a term subsumes all of its descendents, a child may have more than one parent. The key idea of Croft and co-workers has been to use a very simple but efficient subsumption measure. Term x subsumes term y if the following relation holds :

$$P(x|y) > t \text{ and } P(y|x)<P(x|y).$$

Where t is a preset threshold. Thus x subsumes y if documents in which y occurs are a subset or nearly a subset of the documents in which x occurs. The second rule ensures that if both terms occur together more than $t\%$ of the time, the most frequently occurring term will be chosen as the parent. This type of hierarchies seems to be promising.

Using related ideas, [Krishna and Krishnapuram, 2001], propose a framework for modelling asymmetric relations between data. One of the applications of their method is the generation of terms hierarchies similar to Croft and Sanderson ones. [Vinokourov and Girolami, 2000] also propose a probabilistic model with a hierarchical structure for the unsupervised organization of a collection into a hierarchy.

All these recent works rely on the construction of term hierarchies and the classification of documents within these hierarchies. Compared to that, we propose two original contributions in [Njike and Gallinari, 2003]. The first is the extension of these approaches to the construction of real concept hierarchy where concepts are identified by set of keywords and not only by a single term, all concepts being discovered from the corpus. These concepts better reflect the different themes and ideas which appear in documents, they allow for a richer description than single terms. The second contribution is the automatic construction of a hierarchical organization of documents also based on the "specialization/generalization" relation. This allows navigating a collection relying on the subjects appearing in the collection and not only on the terms of the collection.

Automatic generation of non-typed and typed links

In this part, we will introduce the emergent families of ideas relating to the automatic creation of the typed and non-typed links. The non-typed links are the links such as those existing today on Internet. The typed links have more information describing the nature of the link between the documents they link. The fact of typing links can be interesting in more ways: they give the users a navigation context; the types are very useful to target desired information when one does not have time to navigate in all the collection.

The main philosophies of automatic construction of the non-typed links which emerge from the literature are the follows:

- The use of similarity measures by indexing the documents by the terms they contain [Blustein and Webber, 1995] [Green, 1997].
- The use of heuristics [Tebbutt, 1999]
- The reorganization of already linked corpus with the idea that in a good corpus the distance between documents must reflect the power of their similarity [Dean and Henzinger, 1999]

Concerning typed links, although many researchers agree about their importance in hypertext systems as such links might prove useful for providing a navigation context or for improving research engines performances; little work has been dedicated to the automatic methods for the generation of typed links. In the same way, few works were carried out on the useful types of links for the hypertext systems. Some authors have developed link typologies. [Trigg, 1983] proposes a set of useful types for scientific corpora, but many of the types can be adapted to other corpora. [Cleary and Bareiss, 1996] propose a set of types inspired by the conversational theory. These links are usually manually created. The authors propose a semi-automatic technique for creating specialization, detail and example links. Every document is described by a set of attributes or by a set of concepts. This indexation is performed manually. Using this indexation, typed links are deduced using a set of rules also manually developed.

[Allan, 1996] proposes an automatic method for inferring a few typed links (revision, abstract/expansion links). He chose to avoid complex text analysis techniques by deducing the type of a link between two documents by analyzing the similarity graph of their subparts (paragraphs). [Lawrence et al., 1999] automatically generate the "citation" links between scientific articles.

Useful types of link for scientific corpora

As we notice in introduction, the wide availability of the scientific work in electronic form changes radically the way researchers reach the scientific literature. To help the researchers and others users in their mining of this literature, works about scientific corpora organization are necessary. The two types of organization suggested are to generate hyperlinks between documents and to organize documents into concept hierarchies [Lawrence et al., 1999]. However few types of links are proposed within these corpora. The main types of links are "similarities" links where similarities are computed according to several criteria (using the same words, similar headings, and similar citations) and the "citation" links.

What can be the needs of a researcher when navigating scientific corpora?

- To have a global summary of the set of themes present in the corpus
- For a given work, to have close work or alternative sights of this problem (this can be obtain by following the "equivalence" links from the given document)
- To have pointers to the works which are necessary to the comprehension of a paper ("necessary" links)
- To have the history of a methodology: methodology to resolve a problem or types of problem where the methodology has been used to solve them (generation of "methodological" link between two documents using the same method to solve their problems)
- To have references data sets in a field or for some types of problems
- To have some specializations or generalization of a problematic ("specialization" or "generalization" typed links)
- To have some pointers to the articles consolidating or refuting a work ("support" or "refutation" typed links)
- To have some practice cases of the application of a theory ("application" links)
- To have the details or the summary of a work (links typed "summary" or "detail")
- To have the various expansions which were made from an work idea (links typed "future")
- For a given problem, does a solution already exist? ("solution" links)

All these questions suggest different types of links that can exist between the scientific works. The question we tried to answer is to know which are the types of links among those pointed out that can be generated automatically or semi-automatically.

The researcher would also like to know which are the current problems and the emergent communities. The analysis of the links structure [Chakrabarti et al., 1999] [Gibson et al., 1998] [Kumar et al., 1999] of the corpus can help answering these questions.

Models for target types links

For some types of links considered as relevant for the scientific hypertext systems we propose methods or heuristics to generate them automatically. Some methods were taken in existing literature and others are original contributions.

For producing a *global summary of the set of themes* present in the corpus we propose in [Njike and Gallinari, 2003] a method for deriving a hierarchical organization of topics from documents collections. The method automatically derives concept hierarchies from a document collection and automatically generate from that a document hierarchy. The concept hierarchy relies on the discovering of "specialization/generalization" relations between the concepts which appear in the documents of a corpus. Concepts are themselves automatically identified from the set of documents. The proposed method is fully automatic and the hierarchies are directly extracted from the corpus, and could be used for any document collection. Alternatively, this method may be used to create *"specialization/generalization"* links between documents and document parts. It can then be considered as a technique for the automatic creation of specific typed links between information parts. Such typed links have been advocated by different authors as a mean for structuring and navigating collections.

For the *"equivalence" link*, we use the traditional measure of similarity which is the cosine between the vectors representing the documents. The link is generated between two documents if their similarity is higher than a certain threshold (it is a hyper-parameter of the algorithm which can be learned).

The *"summary/detail" link* is induced automatically by the method of [Allan, 1996]. Here a summary is not the same as an abstract in scientific paper. It should rather be seen like a condensed development of a subject, for example the short version of a paper.

The *"citation" link* is induced by the method of [Lawrence et al., 1999]. The "citation" links are very interesting within the framework of scientific work with several reasons: their analysis can reveal some kind of relations between the articles, can identify the significant improvements and criticisms of a previous work, can pay the attention on the corrections or significant retractions on public works, can allow to evaluate the articles, the authors and to analyse the tendencies of research.

For the *"future", "solution", "methodology", "support", "refutation" links* we propose heuristics exploiting the network of citation links between the documents of the corpus for automatically induce these types of links.

The *"future" link*: a document which extends a work of another generally quotes it. The citations network thus enables us to limit our space of research. In addition, we need as an entry of the system a library of sentences reflecting the fact of extending an action. These sentences can be learned on a specific data. The detection of the extending action in the neighbourhood of the citation will allow us to infer a future link between the concerned documents.

The *"solution" link* is based on the same idea with an additional constraint to find within the two documents the same problematic concept and the action of resolution relates to this concept (by using the method of [Njike and Gallinari, 2003] to allow the indexation of the documents by the concepts present in the corpus). Action of resolution is considered to be related to the same concept if the similarity between the resolution paragraph and the concept is superior to a certain threshold, the framework sentences use in this case detect the resolution action.

The *"methodology" link*: methodological papers are much cited, they have a very high degree of citation. The documents sharing the methodological link are detected by the analysis of papers citing methodological paper and their intersections.

The *"support" and "refutation" link*: in addition to cite each other, the documents sharing this type of link must be about the same concepts. For each common concept we analyse the direction of the key words in the documents (if they are employed in a positive or negative way) [Turney and Littman, 2002] and we deduce from that possibly one from the two types of links.

For the last four types of links, heuristics was not tested yet, therefore we do not have an idea of their performances. For the *"necessary" link* we are trying to extend the work of [Morin, 1999] for the detection of the semantic relations between terms to the levels of terms set and documents. This extension might be used for other typed links.

Conclusions and Perspectives

Today the automatic structuring of the collections is a key question. The elements of structure which are the hierarchies of concepts and typed hyperlinks are relevant for this task. Of course all the types of links are not relevant for all the corpora. We have in this article suggested a definition of a good hypertext system to access the scientific literature, in particular the useful types of links for an intelligent access to this literature. We also proposed methods and heuristics for generating automatically this system.

The first results [Njike and Gallinari, 2003] related to the automatic generation of concepts hierarchies which are a good summary of the contents of the collection, the "equivalence" link, the "specialisation/generalisation" link are encouraging and consolidate us in the idea that it is possible to automatically structure the collections, precisely scientific collections. Obviously more experiments should be done and some heuristics remain to test. Generic methods for the classes of the types of links will form part of the next study. Indeed, we can pick up four principal characteristics for the target links:

- links indicating a chronology: future, pre-necessary
- links indicating a reference to an object: citation, data
- links indicating an action of proof: refutation, support, solution
- links based on the similarity: equivalence, specification/generalisation, summary/expansion, alternative sight, pre-necessary

These different categories require various types of modelling: modelling of time, of the reference, of causality action, of the consequence, of conclusion, of consent, or of statistical similarities.

We also plan to use links structure:

- To consider new techniques of visualization of the collection or request results. Concerning the request results, one can use the information of link type in order to gather the documents by categories (summarized, details, pre-necessary…). Concerning a thematic collection for example, one can use the structure of the links in order to produce specific sights to this collection and to propose various modes of navigations in these sights.
- For improving the relevance of research results: the structure of the links can also contribute to improve of the relevance of the results of information retrieval. [Kleinberg, 1998s] shows that the use of the links can help efficiently to find documents of strong authority on a subject (authority), and the documents pointing on document with strong authority on this subject (hub). For a given subject, these two types of documents should be presented to the user. These methods can be improved by refining the concepts of authority and hub with the types of links.

In spite of complements to be brought, we are convinced that the hypertext system we described could be a significant tool in the framework of active research field and would facilitate the access of this information for the researchers.

References

J. Allan. 1996. *Automatic hypertext link typing*. Proceeding of the ACM Hypertext. Washington, DC pp.42-52.

J. Blustein, R. Webber. 1995. *Using LSI to evaluate the quality of hypertext links*. Presented at ACM SIGIR IR and automatic Construction of Hypermedia: a research workshop, Maristella Agosti and James Allan, eds.

Soumen Chakrabarti, Byron Dom, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, Jon M. Kleinberg. 1999. *Mining the link structure of the world wide web*. IEEE Computer 32 (8), 60-67

C. Cleary, R. Bareiss. 1996. *Practical methods for automatically generating typed links*. Hypertext, Washington DC USA

D. R. Cutting, D. R. Karger, J. O. Pedersen, J. W. Tukey. 1992. *Scatter/gather: A cluster-based approach to browsing large document collections*. In ACM SIGIR.

J. Dean, M. Henzinger. 1999. *Finding Related Pages in the World Wide Web*. In Proceedings of WWW-8, the Eighth International World Wide Web Conference

David Gibson, J.M. Kleinberg, P. Raghavan. 1998. *Inferring Web communities from link Topology*. In Hypertext 1998: 225-234

Stephen Green. 1997. *building hypertext links in newspaper articles using semantic similarity.* Proceedings of Third Workshop on Application of Natural Language to Information Systems (NLDB '97)

Jon M. Kleinberg. 1998. *Authoritative sources in hyperlinked environment.* Proc. 9th ACM-SIAM Symposium on Discrete Algorithms. Also appears as IBM Research Report RJ 10076, May 1997

K. Krishna, R. Krishnapuram. 2001. *A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining*. Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management. Atlanta, Georgia, USA. Pp.571-573

Ravi Kumar et al. 1999. *Trawling the web for emerging cyber-communauties*. WWW8 Computer Networks 31 (11-16), 1481-1493.

S. Lawrence, C. Lee Giles, K. Bollacker. 1999. *Digital Libraries and Autonomous Citation Indexing*. IEEE Computer, Volume32, Number6, pp. 67-71.

Mark Sanderson, Bruce Croft. 1999. *Deriving concept hierarchies from text*. In Proceedings ACM SIGIR Conference, 206-213.

Morin Emmanuel . 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse en Informatique, Université de Nantes.

Hermine Njike Fotzo, Patrick Gallinari. 2003. *Génération d'une structure hiérarchique de concepts et de documents à partir de corpus*. Extraction et Gestion des Connaissances RSTI série RIA-ECA volume 17-n°1-2-3. EGC Lyon.

John Tebbutt. 1999. *User evaluation of automatically generated semantic hypertext links in a heavily used procedural manual*. The National Institute of Standards and Technology, Gaithersburg, MD 20899.

Randall Trigg. 1983. *A network-based approach to text handling for the online scientific community*. University of Maryland, Department of Computer Science, Ph.D dissertation.

P.D. Turney, M.L. Littman. 2002. *Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Words Corpus*. National Research Council, Institute for Information Technology, Technical Report ERB-1094.

A. Vinokourov, M. Girolami. 2000. *A Probabilistic Hierarchical Clustering Method for Organizing Collections of Text Documents*. Proceedings of the 15th International Conference on Pattern Recognition (ICPR'2000), Barcelona, Spain. IEEE computer press, vol.2 pp.182-185