Mining for knowledge chunks in a terminology network.

Fidelia Ibekwe-Sanjuan, Eric Sanjuan

8th International ISKO conference, University College London, 13-16 July 2004, 41-47.

IBEKWE-SANJUAN Fidelia [1]
[1] ERSICOM, Université de Lyon 3.
4, cours Albert Thomas
69008 Lyon - FRANCE
ibekwe@univ-lyon3.fr

SANJUAN Eric [2]
[2] IUT de Metz, LITA - Université de Metz
Ile de Saulcy
57047 Metz – FRANCE
eric.sanjuan@iut.univ-metz.fr

Mining for knowledge chunks in a terminology network

Abstract.
This paper examines further a research hypothesis that syntactic variations are an interesting alternative to the clustering approach and they offer meaningful ways of highlighting and organising associated research topics in a corpus. A textmining and topic mapping system, TermWatch, has been developed basing on this hypothesis. Preliminary results obtained on a large IR corpus are promising and call for further systematic investigation.

▪ 1. Introduction.

We developed a textmining and topic mapping system, TermWatch, which clusters text units through prior linguistic processing. The system takes advantage of recent advances in Computational terminology (CT) (Jacquemin, 2001) to enhance the term extraction and variant structuring stage whose output is fed into the clustering scheme. Typically, our end user, a domain expert, wishes to know what topics are dealt with in a huge corpus, what topics are evolving and how each topic is related to one another in other to carry out a competitive intelligence task. S/he needs a global view, a map of the domain research topics embodied in the corpus.

Baeza-Yates & Ribeiro-Neto, (1999) carried out a comprehensive review of clustering techniques applied to the IR field. The basic approach consists in partitioning a collection of documents into many small clusters or groups and then mapping user queries to the most similar clusters. Clustering has also been used to address the specific issue of query expansion in IR systems as well as the presentation of results of a query. Hearst (1999a) reviewed methods of text categorization or of clustering that enhance the presentation of retrieval results. The aim of these studies is not to explain the layout of research topics but to present groups of 'similar' documents in answer to a user's request. Hence, these systems cannot show the user the contents of the documents and how they are interrelated. This information either comes from reading up titles and abstract of publications of these authors (through further search) or from previous background knowledge.

This is precisely the object of the textmining system we developed. According to Hearst (1999b), the goal of textmining is *"to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise*". The most exciting challenge is that of finding, hitherto unknown knowledge from these huge data which will modify the behavior of the user. A remarkable difference between the clustering algorithms used by most IR-clustering and textmining methods and the one used in TermWatch is that the latter does not cluster text units based on co-occurrence. What we use is the types and number of syntactic variation relations between terms. For instance, the terms "*object software*", "*object oriented software*, *objected oriented software testing*" share syntactic relations. The second term is an insertion variant of the first and the last term is a head-expansion of the second term. Here, we will focus on the application of TermWatch to large corpus and show how it

can reveal the structure of a research field as embodied in a corpus of domain-specific texts. The corpus used in this experiment are titles and abstracts of papers published in 16 leading journals in the IR field from 1997-2003. The list is given in the appendix. 3355 titles and abstracts records were thus extracted making up roughly 455 000 words[1].

2. System overview

We outline briefly the system's architecture. We refer the reader to Ibekwe-SanJuan & SanJuan (2004) for a more detailed description.

2.1 Term extraction and term variants search

We extract only terminological noun phrases which are multiword expressions that can appear as compounds (*information retrieval system*) or as syntagmatic NPs with prepositional attachments (*special terminology of information science*). We implemented the term extraction phase as finite state transducers in the INTEX linguistic toolbox (Silberztein, 1993). From the IR corpus, 47 366 term candidates were extracted. In this particular experiment, we chose not to filter the terms at all in a pure textmining tradition, letting the system run in an unsupervised manner. After term extraction, variation identification is performed.

Systems aiming to extract domain terms need to address the variation issue in order to capture the actual state of a domain's terminology. For instance, some of the variants of the term "*academic library*" found in the corpus are : *canadian academic library privilege*, *changing culture in academic library*, *electronic communication in academic library*, *greater utilization of academic library service*, *hellenic academic library link*, *service in malaysian academic library*, *directors of academic library, future of academic library.* Whereas controlled vocabularies (lexicons, thesauri) will tend to retain only the generic term "*academic library*", we retain all its variants found in the corpus. Working directly on variants found in the corpus enables us to discover the other terms which reflect specialisations of the generic concept or associations between the generic concept and its variants. Association is to be understood in its broadest sense here. For instance, there is no apparent semantic relation between "*directors of academic library*" and "*future of academic library*" but both tell us what is being said or done about "*academic libraries*" in the corpus. This is more in keeping with the objective of TermWatch which is to map out the topics contained in a corpus. The subset of variation phenomena currently studied concern syntactic operations namely, expansions and substitutions. Each category is further subdivided along the grammatical axis : variants that affect modifier words in a term, called COMP and those that affected the head word, called CLAS.

COMP relations include left-expansion (L-Exp) : *academic library* → *uk academic library* ; insertion (Ins) : *academic library* → *academic biology library* ; modifier substitution (M-Sub): *academic library users* ↔ *public library users*.
CLAS relations include left-right expansion (LR-Exp) : *academic library* → *canadian academic library privilege* ; right-expansion (R-Exp) : *academic library* → *future of academic library* ; head substitution (H-Sub) : *directors of academic library* ↔ *future of academic library.*
From these relations, a term-variant graph is produced. The next stage is the reduction of this graph through a clustering algorithm.

2.2 Term variant clustering

1Corpus collected from the PASCAL database of the INIST (http://www.inist.fr)

TermWatch implements the CPCL (Classification by Preferential Clustered Link) algorithm defined in Ibekwe-SanJuan (1998). It clusters variants based on a linguistic hypothesis : the relations affecting only the modifier words in a term (COMP) serve as a prior grouping criteria which enables us to gather together terms sharing the same head word and one of the modifier variation relations we specified. For instance the following terms were put into the same component "*information department*, *information science department*, *sheffield university's information department*". To do this, the clustering module computes the connected components on COMP relations. The result of the component building stage is a mono-thematic organization, which is not the desired result. What we seek to highlight is the transversal relation between these lone themes, i.e, what associations have the authors been making between these themes ? To highlight these association, we now cluster the connected components into classes using the second subset of variation relations, the CLAS relations. Let us recall that CLAS relations concern those that involve a shift in the head noun, thus a shift in the topical focus of the noun phrase (NP). Like in most clustering methods, we need to compute a similarity index in order to build clusters. This coefficient is defined as follows :

$$d(i,j) = \sum_{R \in CLAS} \frac{N_R(i,j)}{|R|}$$

where $N_R(i,j)$ denotes the number of R variations between two connected components i and j. A notable difference with other clustering algorithms is that we do not compute this index on the list of terms, but on the set of connected components. The user can set the number of iterations at which the algorithm is stopped and the minimal similarity index to be considered or let the algorithm converge and then choose the results of a given iteration. In this experiment, we chose the results of the second iteration because classes and their layout seemed meaningful. 397 classes of variable sizes were thus obtained containing a total of 1848 terms.

▪ 3. Discovering the structure of research topics

The output is automatically formatted in the GDL (Graph description language) used by AiSee[2] for visualization. We set parameters in Aisee enabling the user to choose which links to view : minimal, weak, medium and strong. Each class can be unfolded to show its internal structure : the components it contains and the most active variants that formed the class. The user can thus perceive the most salient features of a class and immediately judge its coherence from the visualization interface. Due to the huge size of the global image of clusters, it will not be possible to show it entirely here. An interested reader can find more details in Ibekwe-SanJuan & SanJuan (2004). We will focus here on the analysis of a sub network portraying "web" and "internet" related research topics. This is shown in figure 1. 18 clusters were found to be in this network, organised around the core cluster labelled "*testing search engine*" which contains 72 term variants. For instance, the term "*web evaluation*" has as insertion variants "*web search engine evaluation, web site evaluation*, *web search engine evaluation*". Other terms found in this cluster are "*testing search engine*, *alta vista search engine, google search engine, web image search engine*, *web page search*, *query taken from real web search log, search log*". This cluster embodies terms that reflect research on 'web search engine evaluation and comparison'. "*natural language search engine*, *map based search engine*, *intelligent web search engine*" also found in this cluster portray other research trends: the use of computational

2 www.aisee.com

linguistic techniques in enhanced internet search engines or the use of clustering techniques to present results in a graphic form instead of ordered lists of pages.

This intuition remains to be confirmed by a time-series clustering of term variants which will show when the different terms were used in the corpus.

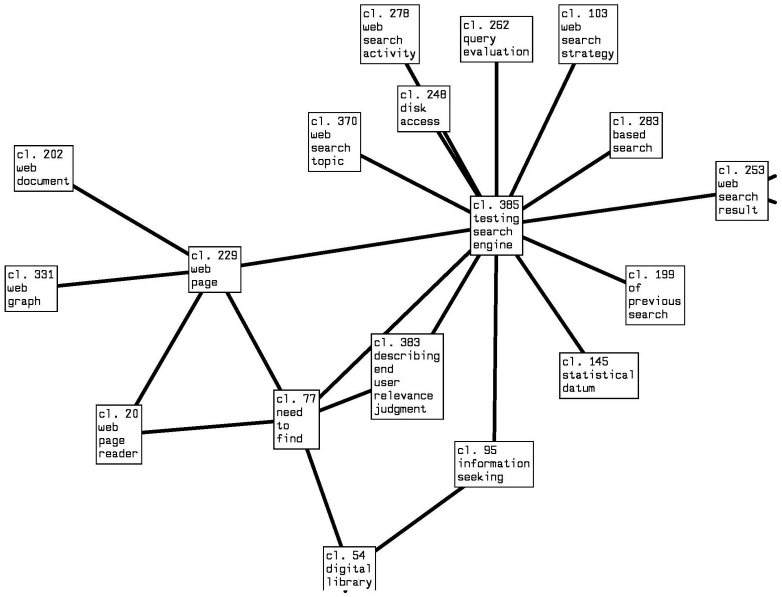

Figure 1. The sub network of clusters around "*Web*" and "*Internet*" research topics.

Surrounding the core cluster are closely related research topics :

1- "*web search activity*" which contains 5 term variants (*search activity question, search question, search activity, web activity, web search activity*).
2- - "*web search strategy*" contains 26 terms reflecting research on 'search strategies' as shown by some of the term variants found in this cluster: *analytical search strategy, high recall search strategy, highly sensitive search strategy, pre stored search strategy, structured search strategy.*
3- "*web search topic*" deals with research on 'search sessions' (*average search session, multitasking search session, image search session, web search session*) and 'search topics' (*search topic, trec search topic, web search topic*).
4- "*web search result*" has 18 term variants such as "*high accuracy search result, refining search result search query result, web search result, search result presentation*" which portray research on different aspects of search results : their relevance (accuracy), their "refining" and "presentation".
5- *'query evaluation'* has 11 terms reflecting precisely this topic : methodologies for query evaluation.
6- *'describing end user relevance judgement*', with 36 terms, reflects research on different aspects of 'relevance feedback' as shown by term variants like "*binary relevance judgment, making information quality judgment, partial relevance judgment, relevance feedback judgment, main evaluation criterion, relevance evaluation criterion, standard web site evaluation criterion.*
7- The link between the core cluster and *'statistical datum'* turned out to be relevant despite appearances. In fact this cluster contains eight terms dealing with the statistical analysis of users searches in digital libraries. The cluster is linked to the central cluster "*testing search engine*" by terms like *"statistical datum from web search log*" thus

reflecting research on the analysis of website log files by search engines in order to improve their performance.

8- Likewise the link *'disk access'* is thematically sound. The latter contain 31 terms that reflect research on distributed database technology used by search engines to speed up the search time. A look at texts associated to the last two clusters confirmed our readings.

The above clusters attest the importance and vitality of research on search strategies, evaluation methods of search engines and performance of search engines.

With the cluster *'information seeking*' (100 variants), we move on to a more theoretic and human-oriented research issue. This cluster portrays research on the information seeking behaviour of different categories of users as shown by term variants like "*human information behavior, cognitive behavior, agricultural scientists information seeking behaviour, lawyers information seeking behavior"* and different aspects of information seeking *: active information seeking, everyday life information seeking, web based information seeking*".

A bit on the outskirt of this subnetwork is a group of clusters formed by "*web page, web page reader, web graph* and *web document*". "*web page*" has 174 term variants and deals with different topics like the terms used in a query expansion task (*query expansion term, search accuracy term, selecting search term*), web pages (*academic web page*), different aspects of "*numbers*" (*closed fuzzy number, key word number, large digital image number*), this last theme being predominant in terms of number of variants and the link between the two topics is not really apparent. At first, the relation between "fuzzy numbers" and web pages were not clear to us but a return to the texts associated to this cluster showed a paper entitled "*Evaluation of SGML-based information through fuzzy techniques*". the abtract showed that the authors were working on knowledge management of web pages based on 'fuzzy grammars'. The cluster "*web page reader*" has 11 variants amongst which we have "*web blind user friendly home page reader*, *web site home page structure'*. The cluster, "*web document*" contains 14 variants dealing with 'electronic documents on the web as shown by the variants "*electronic document interchange, electronic document, original web document, web based electronic document , sgml document*". The three clusters labeled "*based search*, *need to find*, *of previous search*" result from bad morphological analysis performed at the term extraction stage.

The sub-network also highlights the special information resources that should be distinguished from the usual web sites. Indeed, the cluster "*digital library*" connects the above sub-network to another one organized around clusters dealing with "*information service"* and "*public*" and "*academic libraries*".

4. Perspectives

We have applied TermWatch in a totally unsupervised manner to a large text corpus in order to evaluate the capacity of our syntactic variations in forming coherent domain topics and in displaying their organisation. The usefulness of these clusters for textmining or topic mapping tasks is that is enables a domain specialist to visualize in a synthetic way the structure of domain topics from a huge corpus. This is at least time-saving and at best a means of acquiring new knowledge because the knowledge structures exhibited by the clustering algorithm are not necessarily trivial or known beforehand. It is such global and intelligible views that domain experts seek when they are faced with a competitive intelligence task. However, a more systematic evaluation framework has to be set up in order to determine precisely which variation relations are most relevant for the applications considered. We are also investigating (Dowdall et *al.,* 2003) the potentials of the terminology structuring offered by TermWatch in a Question-Answering task (Q-A). Usually, in such a task, if the user's

terms and their semantic variants (synonyms, hypernyms/hyponyms) are not recognised by an IR system, the consequence is usually silence, resulting in failure. The idea is to combine classes produced by TermWatch with a Q-A system in order to "relax" the set of related term variants which the system can consult in order to process the user's query. The utility of these classes for a Q-A task is in capturing more fuzzy relations between terms across different concept families. Another research direction is the integration of semantic variants in the system.

References

Baeza-Yates R., Ribiero-Neto R. (1999). Modern Information Retrieval, ACM Press, Addison-Wesley, 1999, 510p.

Dowdall J., Rinaldi F., Ibekwe-SanJuan F., SanJuan E. (2003). Complex structuring of term variants for Question Answering. Workshop on *Multiword expressions : Analysis, Acquisition and Treatment*. In *41st Meeting of the Association for Computational Linguistics (ACL, 2003)*, Sapporo, Japan, 12 July, 2003, 1-8.

Hearst M. (1999a). *The use of categories and clusters for organizing retrieval results*, *In* Strzalkowski T (eds.), Natural language information retrieval, Kluwer Academic Press, 1999, 333-374. (*Text, Speech and language technology,* vol. 7).

Hearst M.A. (1999b). Untangling Text Data Mining. *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, Maryland, June 20-26, 1999. [Invited paper].

Ibekwe-SanJuan F., SanJuan E. (2003) TermWatch : cartographie de réseaux de termes. *5th Conference on 'Terminologie et Intelligence Artificielle' (TIA'03)*, Strasbourg, 31 March – 1 April 2003, 124-134.

Ibekwe-SanJuan F., SanJuan E. (2004). Mining Textual Data through term variant clustering : the TermWatch system. *Proceedings "RIAO 2004 - Coupling approaches, coupling media and coupling languages for information retrieval*". University of Avignon, France, April 26-28, 2004, 15p. *Forthcoming*.

Ibekwe-SanJuan, F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98),* Brighton UK, 23-28 August 1998, 170-174.

Jacquemin C., (2001). Spotting and discovering terms through Natural Language Processing. MIT Press, 2001, 378p.

Silberztein M. (1993) Dictionnaire électronique et analyse automatique des textes. Le système INTEX. Masson, Paris.

Appendix 1. List of the 16 journals used to constitute the IR corpus

Rank	Nb records	%	rec-cumul	%-cumul	Journal name
1	831	25%	831	25%	Information sciences
2	688	21%	1519	45%	Journal of the American Society for Information Science and Technology
3	283	8%	1802	54%	Information processing & management
4	272	8%	2074	62%	Journal of information science
5	267	8%	2341	70%	Information systems management
6	175	5%	2516	75%	Journal of Documentation
7	176	5%	2692	80%	Information Systems
8	116	3%	2808	84%	Information systems security
9	108	3%	2916	87%	Library & information science research
10	108	3%	3024	90%	Online information review
11	87	3%	3111	93%	Journal of internet cataloging
12	70	2%	3181	95%	Information retrieval & library automation
13	67	2%	3248	97%	Knowledge organization
14	44	1%	3292	98%	Journal of Information Science and Engineering

15	34	1%	3326	99%	International forum on information and documentation
16	29	1%	3355	100%	Information retrieval
	3355	**100%**			