Technical aspects of Thesaurus Construction in TIPS

Jean-Pierre Chevallet

Technical aspects of Thesaurus Construction in TIPS

Jean-Pierre Chevallet

Laboratoire CLIPS-IMAG
385 avenue de la Bibliotheque
B.P. 53 38041 Grenoble Cedex 9, France
Jean-Pierre.Chevallet@imag.fr

Abstract. This paper describes the work done in the TIPS project about the construction of a thesaurus. This construction is a merge from a compilation of data from several web sources. These data comes from manual work, some data are real thesaurus, other are indexing recommendations. The merge is done with automatically extracted terms from large text corpora. The automatic extraction is based on both syntax and statistics. We present in this paper the way thesaurus are built and the results on Scientific corpus in the context of the TIPS project. This short paper emphasis on some technical aspects.

1 Introduction

We present in this document, some technical aspects about the task of building thesaurus for the IRA in TIPS. In the TIPS approach, we have decided not to use thesaurus at indexing time: indexing aspects are not fundamental in this project, and classical single term indexing has been chosen. The TIPS portal proposes a thesaurus allowing to select possible query terms, and also to perceive the domain covered by the indexed corpus by browsing its content. This is possible because a lot of terms are directly extracted from inline document content.

We have built a thesaurus for the TIPS portal by collecting data from actual thesaurus and from other sources like, list of indexing terms or structured directory for document classification. Before going into details of the thesaurus construction in TIPS, we just mention some general facts about thesaurus construction.

Manual thesaurus building is a hard task but in this way, one can guarantee a good quality of the collected terms. So we can present these data to the end user for browsing. Maintaining such a thesaurus up to date is also costly. On the other hand, automatic Thesaurus building is quite human costless but the quality is not guaranteed. It relies on the content of document sources and also on the Natural Language treatment implemented. Our goal in this project is to combine both approaches. We will compile manually-built data, and extract terminological knowledge from documents, and finally merge these two sets into a final structure that will be proposed for browsing. Our building steps are then the following:

Extract a terminological base from documents by means of automatic full text analysis;
Compile existing accessible thesaurus and terminological sources;
Validate and filtering automatically obtained terms by confrontation with manual thesaurus and by limited manual inspection;
Merge both data sources. In this step, one can propagate some information from manual thesaurus to automatic thesaurus like the known domain of a term.
Structuring the term set using and propagating extracted links from existing thesaurus, and by the computation of syntax variations.
Integrate the final thesaurus into TIPS portal through the Information Retrieval Assistant.

In the next section we go through these steps in detail.

2 Manual thesaurus construction

The goal of this part is to format a set of data related to the domain covered by the documents stored in arXiv base that is used in this project. We have to first find data sources available on the Web, after to select and process sources that are in the closeted domain of arXiv. The final selected sources are describes at the end of this document. This part of the process is done in the following steps:

Identification of the data sources. It is done by a manual search on the web using web retrieval engine.
Download the data so they can be locally processed.
Inspection of the raw data sources. This step aims to understand the content of the source, to perceive the logical structure in order to generalize rules for extraction.
Extraction rules are then set up according to the relation we have chosen.
Implantation of the rule in a dedicated software. We have used Perl script for this task.
Production of the extracted data an final manual inspection for an estimate of the quality of the result thesaurus.

We now present the automatic term extraction.

3 Automatic thesaurus construction

Thesaurus is extracted from full text by means of syntax analysis. In this part we detail terms extraction and structuring steps that define the automatic thesaurus construction. In this thesaurus, we have a generic relation based on syntactic variation. We also obtain a non typed relation (a sort of "see also") based on conditional concurrence probability which are known as Knowledge Discovery in Text techniques (KDT) [2]. We will not develop this aspect in this article has we don't not have yet the results.

We have used our IOTA system for all tasks except the first one: the full corpus tagging using a part of speech tagger. We have used the Brill tagger [1] because our IOTA system accepts only French texts as input. Thus we have had to develop a coder from Brill tagger to our IOTA format in order to use the rest of our system for all of the other text treatments. For the whole process, we have the following steps:

Transfer of all documents from ArXiv base via simple ftp protocol.
Formatting documents: some must be uncompressed. We have only treated documents in Latex format [1]. This treatment aims to extract text, to split them in sentence, and to adapt some typo in order to be accepted by the English tagger (see next part). Only full text is kept: all bibliography, figures, tables are discard, and also all part before and after de document Latex tag.
Tagging is done using the Brill tagger. The raw output must finally be converted in a format compatible with our IOTA tools.
Term extraction aims to extract all noun phrases from the whole corpus.
Term structuring is the last step that computes a tree of composed terms.

In the two next sections, we develop term extraction and structuring.

3.1 Term extraction

Term extraction is based on part of speech templates. These templates are used to extract noun phrases. In English as in French, most of these phrases are about 2 or 3 full words long. Full words are nouns or adjectives. Longer terms are less frequent and are less numerous. It is useful to extract longer terms if we take into account term variation. If not, we then have two different terms that are synonym in the sentence context. In fact long terms (noun phrases) usually appears once and rather at the beginning of texts. Shorter version then appears in texts as variations of longer terms.

Knowing this linguistic fact, it seems then important to compute co-references between terms and also between terms and pronouns. In TIPS, we do not compute these co-reference paths. Our goal is only to extract and structure terms from the all corpus. Moreover ambiguity between two term variations is very rare because size length is a sort of guaranty against term homonymy. Hence we promote the resolution of term variation and then the co-reference phenomenon, not at the sentence level, but at the end of extraction, so at the corpus level. This approach enables us to use frequency term information to choose the right term variation. This is the next treatment detailed in the next part. This choice explains why we extracted full size terms and so why we do not limit ourself to 2 of 3 terms length. Here are some examples of extracted phrases related to the word "algorithm":

```
randomized bidding algorithm ADJQ SUBC SUBC
optimal randomized bidding algorithm ADJQ ADJQ SUBC SUBC
```

[1] Using detex : http://www.cs.purdue.edu/homes/trinkle/detex/

```
pseudopolynomial time algorithm ADJQ SUBC SUBC
forward search algorithm ADJQ SUBC SUBC
algorithm for matrix multiplication SUBC PREP SUBC SUBC
simple dynamic programming algorithm ADJQ ADJQ SUBC SUBC
cubic time algorithm ADJQ SUBC SUBC
iterative algorithm ADJQ SUBC
simple polynomial time algorithm ADJQ SUBC SUBC SUBC
algorithm for query evaluation SUBC PREP SUBC SUBC
```

Terms are followed by the corresponding part of speech. In the next section we present the structuring of this set of terms that leads to the thesaurus.

3.2 Term structuring by means of syntax

We used two sorts of term structuring. One is based on syntax and cover the term variation phenomenon, the other is based on global document term concurrence and expresses a more general sort of term relation. There are some attempts to automatically acquire from text a given type of relation, like hyponyms [5]. Some other approaches uses context defined by syntax [3, 4]. The Sextant system, uses syntax dependences between noun/noun, noun/verb, and noun/adjective. The underlying hypothesis used is that terms sharing contextual dependencies are semantically related. This approach is not able to qualify the extracted relation. Other systems like Xtract [6] are only based on co-occurrence statistics computed into a five word windows.

For this project we have chosen the combination of two methods : one based on syntax and term variation, combined with one based on term co-occurrence in document using dependence probability.

The syntax driven structuring deals with the all set of full length extracted terms from the all corpus. The system tries to link terms using variation rules. A variation rule is a couple of two part of speech patterns. The left pattern is the trigger of the rule. A rule is fired if the input term matches the part of speech tag sequence of the pattern. The right pattern is the production part. It produces a shorter term by reordering and reducing the set of tags of the right pattern. Applying a rule produces a short reordered term. The goal of such a rule is to link two term variations: a larger and a smaller variation of terms. Here are some examples of such rules. For each rule, one have an example of derivation and the rule itself.

```
deterministic algorithm -> algorithm
ADJQ SUBC <VGEN> 2 .
```

This rule expresses the variation from a term without the adjective that qualify the substantive. The right part of the rule is a sequence of part of speech. The left part is the sequence of word that are kept for the associated term. In this rule, we only keep the second word of the term.

```
positive acceptance probability -> positive probability
ADJQ SUBC SUBC <VGEN> 1 3 .
```

This rule illustrates a term variation by insertion of substantive.

```
probability distributions for sequences of every finite length
  -> probability distributions
SUBC SUBC PREP SUBC PREP PREP ADJQ SUBC <VGEN> 1 2 .
```

This last example, shows a term split at a preposition.

In order to avoid combinatory explosion and production of meaningless terms, the system only attempts to link actually existing corpus extracted terms. Hence, in this approach we have to first extract all possible terms from all documents before the application of these rules. All these rules have been proposed if we have at least one good example of term variation. A set of derivation rule is then language dependent. Here is an example of linked terms produced in this way.

```
optimal randomized bidding algorithm for the case of multiple bidders
-> optimal randomized bidding algorithm
-> randomized bidding algorithm
-> bidding algorithm
-> algorithm

known optimal algorithm
-> optimal algorithm
-> algorithms
```

In case of two rules that can be fired simultaneously, we have a preference for the one producing the most frequent term. If both possible terms have the same frequency in text, we produce them both.

4 Results

In this part, we present some information about thesaurus and documents that have been treated in this project. First the list of available online thesaurus that have been treated and merged and then, some data about documents that have been analysed.

4.1 Data source list

We have treated a list of seven thesaurus. These thesaurus have been chosen because they are related to domains that are present in the ArXiv document base, and second, because there where available on the web. We have extracted from them four relation types:

Generic is hierarchic relation. A term a is a generic of a term b if the meaning of a includes the meaning of b. Hence b is a specific term of a.

Synonym is used when a term a can be used in place of a term b.

Context is a relation that express that a term can be used in the context of an other term.

See is a general relation without a precise meaning. It is often called "seealso".

The table 1 sum up the results. We have found very few synonyms : in only one thesaurus. The context relation is also not very frequent (two thesaurus). The more common relation is generic and after the "see also". The set of term after merging is 13 809. Only 5% of terms are common. Finally, we have obtain an average of 2 relations per terms, which not very important.

Here is the list of thesaurus and other data sources treated:

aao This Astronomy thesaurus is very important. It is composed of 2 846 terms. (http://darmstadt.gmd.de/ lutes/thesalpha.html)

arxiv ArXiv is the organization of the document repository. It is not really a thesaurus but rather a classification scheme for clustering documents in the base. We used 440 terms. The directory structure consists in a whole of chapter and sub-chapter. We deduced from this a set of relations between each chapter and its own sub-chapters. These relations are all GENERIC relations and they were built on the following scheme (for a given chapter):

```
chapter GENERIC sub-chapter 1
chapter GENERIC sub-chapter 2
```

We used a perl script to realize this transformation. The data are reachable at: http://arxiv.org/archive/)

jhep is a very short list of terms on High Energy Physics. This file as the ArXiv data can be seen as a set of chapter and its sub-chapter. The difference with the ArXiv directories is that it also contains sub-sub-chapter but the treatment is exactly the same. We create GENERIC relations between a chapter and its sub-chapter and recursively between a sub-chapter and its sub-sub-chapter. The data are at: (http://jhep.sissa.it/JOURNAL/keywords.html)

msc MCS is a thesaurus dedicated to mathematics. It is structured in three sub levels. This thesaurus is a little more complicated in its structure than the precedent thesauri. As the arxiv and jhep data the msc thesaurus has a chapter hierarchy but it contains also explicit relations SEEALSO and other relations as For ... see, etc and e.g. . In a first stage, we extract from this thesaurus the simple chapter hierarchy and we thus deduced the same GENERIC relations as for arxiv and jhep thesauri. In a second stage we treat the other relations of the thesaurus.

We explain below how we treat each of these relations with an example for each initial relation type. In fact we follows the links in the data. Each entry is identified by a symbolic value.

```
13J30 Real algebra [See also 12D15, 14Pxx]
```

From this sentences we deduced 2 SEEALSO relations:

```
real algebra SEEALSO fields related with sums of squares (12D15)
real algebra SEEALSO real algebraic and real analytic geometry (14Pxx)

14D20 Algebraic moduli problems, moduli of vector bundles
{For analytic moduli problems, see 32G13}
```

From this sentence with also extracted 2 relations:

```
algebraic moduli problems, moduli of vector bundles
  GENERIC analytic moduli problems
analytic moduli problems
  SEEALSO Analytic moduli problems (32G13)

03F45 Provability logics and related algebras (
      e.g., diagonalizable algebras)
```

We decide to understand examples as specific concepts of the thesaurus entry. We have chosen not to take into account this information detail. Thus we have :

```
Provability logics and related algebras
  GENERIC diagonalizable algebras
```

In the same ideas, we have includes these sort of information into the generic specific relation. The reduced set of relation type induces this drastic choice. We have preferred to include the most information as possible in consistent way.

```
12D15 Fields related with sums of squares
  (formally real fields, Pythagorean fields, etc.)
```

We obtain :

```
fields related with sums of squares
  GENERIC formally real fields
fields related with sums of squares
  GENERIC  pythagorean fields
```

The final thesaurus is made up of the relations extracted in the first stage and those extracted in the second stage. The original file is : http://www.ams.org/msc/

pacs PACS thesaurus is about physic and astronomy. It contains 4324 terms related to condensed matter physics, material science and microelectronics. We only used these sections of the PACS thesaurus. This thesaurus is very similar in his structure as the msc thesaurus. Only some elements of syntax are different. We thus treat it in the same way as the msc thesaurus. (http://www.aip.org/pubservs/pacs.html)

schlagw The SCHLAGW thesaurus is more a list of recommended indexing terms than a real thesaurus. We have extracted 1552 terms from it. (http://www-library.desy.de/schlagw.txt)

spires SPIRES is an important thesaurus. We used only the physics part. This last thesaurus is in a simple form. It contains 2 types of relations: see and see also. For the see also relation we only modified the form of the relation. For the see relation we deduced a :
GENERIC relation, if the first term is included in the second
SPECIFIC relation, if the second term is included in the first
SEEALSO relation in the other cases
The data are found at:
(http://www.slac.stanford.edu/spires)

We sum up in table 1 some figures about the treatment of the manual sources.

Table 1. Treated thesaurus

Thesaurus	Theme	See	Generic	Context	Syn	relation	term
AAO	astronomy	8 111	2 432	429	0	11 972	2 846
ARXIV	high energy physic	0	440	0	0	440	115
JHEP	high energy physic	0	124	0	0	124	126
MSC	mathematiques	1 450	4 971	0	0	6 421	4 810
PACS	astronomy, physic	488	3 836	0	0	4 324	3 912
SCHLAGW	physic	1 228	964	186	64	2 142	1572
SPIRES	physic	1 198	343	0	0	1 541	1 191
Total		12 475	12 810	615	64	26 964	14 572
Total	After Merging					26 964	13 809

4.2 About the analyzed documents

We have analyzed quite all content of the ArXiv content. This base has been indexed for the TIPS portal demonstration. We have analysed about 300,000 English documents in latex format. All theses documents are splited into 40 categories and sub categories (see the ArXiv thesaurus above). In the table 4.2 we present some figures obtained on some of them. This table shows the following information:

Doc nb is the number of treated documents in the sub category. We can notice that some categories have very few documents compared to others.
Voc size is the number of single terms found in the collection of documents. We can notice this number is not directly related to the number of documents.

Term nb is the total number of full length terms found. We can see the impressive amount of different terms found. These figures show that word combination produces between 5 and 6 times composed terms more than single terms. In fact it is not such a quantity as we know lot of terms are more than 3 words long.

Hapax is the ratio of terms that appears only once in the corpus. This figure is important because we notice that for every corpuses this value is stable. About 80% of composed terms appear only once !

Max frequency is the maximum frequency of terms. It means the maximum number of documents in which a term can appear. This value is always 3 or 4 times less than the number of document. This value is interesting because we can suspect a term to be useless if it appears in too many documents.

Variation is the number of relations that has been computed. Generally speaking, we notice that we do not have found a lot of relations regarding the number of terms extracted. This is probably due to a reduce set of rules. Theses rules have been set up in incremental and manual ways. We do not know exactly how many rules are useful to cover the maximum of interesting term variations.

Relation is the number of terms that are found in the variation relation. Again we note an important loss of terms due probably to a lack of relation rules.

Table 2. Analyzed documents

Theme	doc Nb	voc size	term nb	hapax	max freq	variation	relation
acc-phys	71	5 578	4 912	87.0 %	12	471	609
adapt-org	781	19 729	46 399	85.2 %	171	9 912	11 881
alg-geom	1 913	34 442	103 669	81.0 %	989	22 379	26 467
astro-ph	18 051	232 567	1 234 090	83.0 %	4 014	273 747	315 298
chao-dyn	3 762	58 528	257 239	83.7 %	1 150	59 364	68 624
cond-mat	62 973	388 581	3 426 576	83.0 %	21 568	618 998	712 604
computer	2 500	53 103	158 570	83.0 %	354	4 177	46 525
hep-ph	63 703	323 485	1 851 639	80.8 %	13 747	350 261	403 059
math	28 444	276 423	1 198 857	80.5 %	27 700	233 887	270 127

5 Conclusion

We have built for this project an important thesaurus related mainly to physics, astronomy and mathematics. We have produced a very huge amount of terms from the available scientific article of the ArXiv pre-print document base. Hence, we have proven that it is possible and useful to run some simple Natural Language techniques in order to automatically built a very important collection of

terms in an automated way. The resulting user interface has not been tested with real users yet. The test done was only on a small set of terms. So we do not know at this moment, the pros and cons brought by the capacity of browsing throught such a huge base of terms.

I thank Carole Bergamini for her help in extracting data from existing thesaurus and launching the processing of the latex file for the construction of the automatic built thesaurus. I thank also Christophe Hoang for the development of the part of the IOTA system that computes the syntactic variation and for the code that merge all data into one unique base of term and relation.

References

[1] Eric Brill. English tagger. In *http://www.cs.jhu.edu/ brill/*.

[2] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In *Proceeding of the First International Conference on Knowledge Discovery KDD'95*, pages 112–117, August 1995.

[3] Gregory Grefenstette. Use of syntactic context to produce term association list for text retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM press Copenhagen, Denmark*, pages 89– 97, 1992.

[4] Gregory Grefenstette. Automatic thesaurus generation from raw text using knowledge-poop techniques. In *Making sense of Words 9th annual Conference of the University of Waterloo Centre for the Oxford English Dictionary and Text Research, Cambridge*, pages –, September 1993.

[5] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France*, July 1992.

[6] F. Smadja. Retrieving collocation from text : Xtract. In *Computational Linguistics*, pages 143– 177, 19(1) 1993.