Product taxonomy and web directory as support for ontology engineers

Jan Nemrava NEMRAVA@VSE.CZ
University of Economics, Prague, W.Churchill Sq. 4, 130 67 Prague 3, Czech Republic

Abstract

This paper presents our attempt to build a text mining tool for collecting specific words – verbs in our case – that usually occur together with particular product category. These verbs could be used as a support for ontology designers in creating relations between entities and concepts. As the ontologies are headstone for the success of the semantic web, our effort is focused on building small and specialized ontologies concerning one product category and describing its frequent relations in common text. We describe the way we use web directories to obtain suitable information about the products from UNSPSC taxonomy and we propose the method how the extracted information could be further processed.

1. Introduction

It is generally known that Semantic Web is intended as "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al., 2002). Most of current web is easily human-readable, but the amount of information provided is becoming human-unprocessable. This is the main reason for developing viable and usable Semantic Web, which allows expression of the semantics of the data precisely enough so that it is machine interpretable. Semantic markup of the data is achieved through the use of ontologies, because they have shown to be right answer by providing a formal conceptualization of a particular domain shared by group of people. As the Semantic Web heavily relies on ontologies they must be reliable and complete. Manual population and construction still remains tedious and error prone and this could result in knowledge acquisition bottleneck (Maedche, 2002). Some semi-automatic or automatic methods are needed and a lot of scientists are putting their efforts in research on this field. In this paper we use HTML pages restricted to specific domain as source of relevant text for semi-automatic extraction of verbs (in (Kavalec & Svátek, 2002) identified as the most representative type of "indicator terms") as relations that usually occur together with particular product category. These relations could be used for construction small terminological ontologies attached to specific types of products and services. As analyzing of large amount of knowledge-sparse text with full linguistic analysis would be too demanding, shallow linguistic methods, typically relying on POS tagging and/or shallow parsing, were chosen. The "side-product" is that extracted verbs could be further used for information extraction.

No doubt that unstructured data (HTML pages in our case) are useful source of information and fulltext search engines are useful tool for retrieving information, but using just fulltext search without any restriction to domain would lead to ambiguous and confusing results. According to our opinion the proper kind of domain restriction could be introduced by using web directory DMOZ, as world most accepted directory, where solid degree of the validity of information is ensured by thousands of volunteers who are responsible for the content. On the other hand, we have to consider that it is known that web directories are generally not valid taxonomies and this impede using them as ontologies (as discussed later). These are main reasons for considering some valid taxonomy that meets some of the requirements that ontologies have – at least that subheadings are specialization of headings. For this reason The United Nations Standard Products and Services Code[2] (UNSPSC) seems to be suitable choice. In this paper we discuss the possibility of mapping DMOZ to UNSPSC and we present attempt to obtain verbs that would be used by ontology engineer as support for creating relations in ontologies.

In this paper, we first describe the reason why UNSPSC was chosen, and why we use directories as source for our data. In second section we introduce our method to identify verbs related to products and in third section we describe experiments and the results. At the end of the paper related work and our future plans are discussed

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[2] http://www.unspsc.org

2. Proposed method description

2.1 DMOZ and UNSPSC

The reason behind using UNSPSC is that we would like to join this taxonomy and its list of product categories with content of company websites to gain valuable information about verbs that usually occur in one sentence with some product category from the taxonomy. For this purpose we build a tool that collects these verbs from given web pages. Presented text mining tool is based on combination of catalogue and fulltext search engine. Our approach exploits redundancy of data on large data repositories like World Wide Web. We are using the knowledge stored in hand classified web directories, DMOZ in our case, and we use their ability to provide web sites relevant to term we have chosen. The problem is that the entry website page (i.e. the one that is linked from directories) usually does not contain much or even any text. When it does, it hardly ever describes the product or the offered services. This led us to use the fulltext search engines with restriction to particular website to ensure that we discover all term occurrences in content of whole company's website.

UNSPSC is freely available in standard ontology format from Protégé[3] website, it contains 16.000 unique products and has unambiguous structure, it could be suitable for use in this field. Each of these 16 000 terms describes particular product category which can hardly be further divided. On the other hand this raises problem that UNSPSC tree leaves (product categories) significantly vary from the directory headings in commonly used directory structure including DMOZ directory. At current time there aren't any tools to automate the process of assigning right UNSPSC category to relevant DMOZ category so it must be done manually by choosing product from taxonomy and then finding appropriate category in directory. There are either a lot of categories describing desired product category or there are none. In the first case (i.e. more categories are found for one product name), we focus on Business branch where we expect manufacturers and companies to have listed their products-offering websites. The latter case – where no category is found – is worse and we have to find similar category, or find similar product. These two issues disable this part of our work to be done automatically. Next part discusses possible techniques how to find appropriate DMOZ directory for UNSPSC product category and by that to obtain relevant data that are essential for the tool to work. We have three different methods how to deal with this problem.

[3] http://protege.stanford.edu/

2.2 Finding UNSPSC leaves in DMOZ directory

Two problems are common for all methods. They are the level of ambiguity and the amount of websites that are in directories. The most ambiguous are one-word product categories, but they mostly have corresponding category in DMOZ. While when searching for multi-word product categories it won't be so easy to find appropriate category but we won't face the problem of ambiguity.

Three different scope of search could be used to get proper list of web sites. One tool enables us to list all categories containing desired product category. Using specific restrictions and regular expression in a search box it is possible to get a list of very relevant nodes. Second tool could be used to search in websites description provided by DMOZ editors, as they sometimes contain the main products names offered by some manufacturer. Advantage of this tool is that it could be further restricted to particular branch (e.g. Business) to avoid possible ambiguity on non-manufacturer's web sites. The lowest level of generalization is using fulltext search engine. If none of previous two methods provides successful results, the product category may be so specific, that fulltext search is only one way to gain relevant web sites.

These methods should ensure enough relevant data for further processing. The question is whether it would be possible to automate the process of assigning proper DMOZ directory to UNSPSC product category. Automation process is made worse by some generalized names of categories, e.g. *Hotels_And_Motels_And_Inns*. It would be possible to use this word separately but when we consider category *Hotels_And_Lodging_And_Meeting _Facilities* the separation wouldn't probably be useful enough.

2.3 Obtaining verbs from relevant web sites

As stated in the introduction we would like to obtain so called „indicator verbs" that characterize particular term (product category in our case) in UNSPSC. In our test we found 7 nodes in DMOZ corresponding to the same number of products from UNSPSC from "Material handling" field. This sample of products categories was chosen because each of these product categories in UNSPSC had exactly the same directory name in DMOZ. We are aware that this small number limits the representativeness of results, but as stated above, finding proper nodes is a complex task. Only several common verbs were obtained and they had to be classified manually, as we don't have any other categories to be compared with results from this. Next paragraph describes the text mining tool that collects data from a selected directory category.

Table 1 depicts subtasks of the tool. The input data for this tool are the *URL of directory in DMOZ* as list of relevant websites to chosen term and the *name of product category* chosen from UNSPSC. The first part uses *link*

extractor to obtain all company's web sites URLs. The list of extracted links is stored in file for further processing. In case no proper DMOZ directory had been found and the data were obtained by another level of directory search, it is possible to start the script with a list of URLs. Every URL from the list is then inserted into Yahoo Search Engine with the term we are currently exploiting and the parameter "site" is added. This ensures that the particular term is only searched on the selected web site. This process is repeated until all URLs from the list have been processed. There is a limitation of fulltext search engine which allows only limited amount of queries in a certain time period. We only store first 10 links from every domain, but it is only matter of setting of script and here we see a possibility of extracting more data. Up to 100 links from every company URL can be stored.

Table 1. Task sequence decomposition

1) Input: URL of DMOZ directory containing companies that manufacture desired product.
Output: List of URL of companies.
2) Input: URL of company website
Output: List of web pages containing the target term.
3) Input: Web page containing the term
Output: File with extracted sentences containing the term
4) Input: Sentence with term.
Output: Extracted verb.

List of several hundreds of URLs (depending on number of links on the list), where our desired term occurs in the page full text, is a source for next step. Next task is to extract every sentence from this set of links where the selected product category occurs. The task is carried out by means of regular expressions and finding occurrence of the term in a set of documents. As the sentences are discovered and saved into file we need to carry out some syntactical analysis to discriminate verbs from other lexical units. It is done by Adwait Ratnaparkhi's Java based Maximum Entropy POS Tagger (MXPOST) (Ratnaparkhi, 2004). The extracted verbs are then compared with each other to find similar verbs, and number of occurrences is counted. The verb similarity is based on simple character comparison, which groups together verbs with various grammatical forms. No semantic-based comparison had been implemented. Using WordNet[4] database and its ability to discover word stem from any word form we assure that part-of-speech tagger didn't committed mistake during tagging the verbs. If mistakes were made, WordNet discovers them and it also provides lemma for each inserted word, which makes storing of verbs much easier.

We store all extracted verbs as matrix in relational database, where the discovered verbs represents rows and the desired terms represents columns. The intersection of row and column is the number of verb occurrences in all web pages of the directory category. We can use these sets of verbs and number of their occurrences for further examination on how the verbs characterize some broader term and if a small ontology could be build for every term. The goal of effort is to build separate lists of verbs such that some only characterize specific product types, while some other characterize whole product areas. We started with node *Handling materials and products* from this category.

3. Experiments

In extracting selected information we currently face the problem of how to set proper amount of web sites to be analyzed and what is the right amount of words taken in the surroundings of the word. As stated above in this particular experiment, 10 web pages from one company website and one sentence from each web page containing examined word is taken. Our tests showed that this can be sufficient amount of data for extraction of verbs for some common product types, but other product types from the same category of UNSPSC suffer from lack of data for extraction because they don't have appropriate category (node) in DMOZ.

Using the above described tool we have built a database containing 303 verbs for 7 product categories from *handling material* category. These are only words that have appropriate category in DMOZ and therefore our approach could be used for their extraction. These verbs occurred 7300 times near the selected terms.

Our goal is to find some method that would enable us to categorize verbs as either:

- *Common* for most products.
- *Characterizing* one branch of products
- *Specific* for small group of products, or even only *one product.*

Even from seven product categories – as expected – some verbs are obvious to be entirely neutral and do not characterize the products at all. According to three methods described later, verbs *be, have, provide* and *use* are common for all sentences describing any product. Then there are verbs describing activities connected with manufacturing of any types of products e.g. *design, require, offer, make, contact, manufacture, develop, supply,* etc. More specific for our branch might be verbs

[4] http://wordnet.princeton.edu/

describing activities related to manipulating with material. They are *handle, lift, install* and *move.*

We experimented with three different measures that could separate specific verbs from more general ones. First and second are normalizations of frequencies to eliminate the influence of very frequent verbs. Normalization based on proportions of product categories in collection is the first, **Croft's normalization** using elimination of high-frequency terms with a specific constant is the second and **TF/IDF (**Salton & Buckley, 1988), which relies on indirect relation between verb occurrences and its importance for product category, is the last. We also tried **Lift measure (**Brin et al., 1997) but it didn't provide satisfactory results for aggregate values. We plan to use it for individual product category in the future as it measures how many times one verb occurs more often with one term together than expected if they were statistically independent.

$$F_{ij} = f_{ij} * (V_{tj} / V) . \quad (1)$$

We tried these three methods to class verbs to their corresponding groups of verbs. All methods provided quite similar results. The first is *normalization* described by formula (1), where *Fij* is normalized frequency, f_{ij} is the frequency of verb *j* in product category *i*, V_{tj} is sum of all occurrences of product category *i* in collection and V is total number of collected verbs. Then V_{tj} / V represents how many per cent has product category *i* in collection. We recalculate whole matrix to get numbers ranging from 0 to 43 representing the normalized frequencies showing that the verbs with high value (30-43 in our case) are independent on the product category and thus they can be considered as common one. Verbs with values from 10 to 30 are not so often and they could be used as branch descriptors. The rest are with frequency lower than 10 are out of our interest for this moment.

$$cf = K + (1 - K) * f_{ij} / m_{ij} . \quad (2)$$

Croft's normalization (2) moderates the effect of high-frequency verbs, where cf_{ij} is Croft's normalized frequency, f_{ij} is the frequency of verb *j* in product category *i*, m_i is the maximum frequency of any verb in product category *i*, K is a constant between 0 and 1 that is adjusted for the collection. K should be set to higher value (higher than 0.5) for collections with short documents. We used 0,3 as there are no different between 0.3 and 0.5 in our table. With this formula we get sum values for every verb ranging from 2.1 (7 product category × 0.3 for zero occurrences) for no occurrences of verb in our database to 8.58 for the most often verbs. Verbs with number above 5 normalized occurrences are significant for us as the common indicator while verbs between 3 and 5 normalized occurrences could be taken as the products representing verbs. The rest, with 3 and lower occurrences is for us as in previous method uninteresting.

$$w_{ij} = f_{ij} * \log_2(N / n) \quad (3)$$

TF/IDF (term frequency / inverse document frequency) (3), where w_{ij} is a weight of verb in product category *i*, f_{ij} is the frequency of verb *j* in product category *i*, *N* is number of all verbs in collection and *n* is sum of verb *j* occurring in all product categories. TF/IDF is technique that gives verb a high rank in a document if the verb appears frequently in a document or the verb does not appear frequently in other product categories. In other words, a verb that occurs in a few product categories is likely to be a better discriminator than a verb that appears in most or all categories. As a result in this test we got values from 0 to 1350. Where as usual, the highest values between 1000 and 1350 are verbs that occur independently on selected product category and we consider them as common verbs. We are much more interested in verbs with value starting around 300 and ending at 1000. As stated above, these could be used as identifiers of the product category.

3.1 Evaluation

In our trial we only examined 7 product categories from one UNSPSC node and hence we are not able to classify verbs into four categories as we suggested in part 1. We only classified them on common and specialized verbs. The first 15 results with values from each of described method are shown in Table 2.

Table 2. Comparison of three methods

	lemma	Per cent	lemma	croft	lemma	TFIDF
1	have	43,01	have	8,58	have	1 318,40
2	provide	40,38	provide	7,41	provide	1 164,76
3	design	39,36	design	7,14	design	1 119,10
4	use	37,29	use	6,38	use	1 028,17
5	lift	26,47	require	5,32	require	802,81
6	require	26,43	handle	4,70	lift	703,11
7	handle	19,81	lift	4,70	handle	676,10
8	mount	17,75	offer	4,68	offer	648,62
9	operate	17,66	allow	4,31	allow	596,96
10	truck	17,61	include	4,30	contact	587,38
11	allow	17,25	please	4,29	move	582,57
12	contact	16,37	make	4,18	please	582,57
13	offer	15,99	contact	4,15	include	572,89
14	meet	15,91	need	4,06	meet	538,52
15	include	15,49	install	4,06	make	538,52

The table could be divided into three layers according to proposed verbs categorization. It can be observed that the most common verbs are in first (four) rows, and those more specific for tested category are beneath them (e.g. row 5-10). The remainder is not specific enough to be considered as important for these products.

To verify possible general use of this tool it would be necessary to test some other nodes and product categories. So far we have encountered some categories that were very hard to process because their web sites contained very little free text that could be used for the tool. These were mostly presentations of end-user products whereas product categories for B2B commerce are more informational than presentational. Third group of products is created by some categories that are divided in more formal than – more commonly used – logical way. This is for example division of *hotels* to *double rooms* and *single bed rooms*.

As already mentioned the reason why this approach cannot be automatically run are mainly the non-corresponding items from product taxonomy to categories in widely-spread product catalogues. Our plans and intentions for the development of this tool are stated in future work section.

4. Related Work

The idea of combination information extraction with ontology learning has been described by Maedche in (Maedche & Neumann & Staab, 2002). Idea of using identified words to extract more words was in (Riloff & Jones, 1999) called *mutual bootstrapping*. This paper follows up with work (Kavalec & Svatek, 2002) as it brought to this field use of web directories. While Brin (Brin, 1998) uses fulltext search engines to obtain data from arbitrary sources we only use search engines for obtaining full text from the websites we have previously identified by another method, because our data are less structured and can be mistaken easily by ambiguous meanings of terms.

5. Future Work

As described in this paper, there are currently some limitations of this approach; they are mainly caused by lack of data to be mined from websites for specialized terms. We proposed some techniques how to overcome these limitations. One of them is relaxing restrictions of fulltext search engines and the second is searching in all subdirectories for given terms in all whole DMOZ branch tree structure. All our plans for future stem from our effort to obtain as much data as possible and also better automation of whole process. As soon as we obtain verbs for more branches we could try to classify the verbs into four categories as proposed in section 3 and use them for creating ontologies with relations labeled with extracted verbs. This technique doesn't have to apply on verbs only as it is possible to extract any other word classes from the extracted sentences (e.g. adjectives to construct entity properties or nouns).

6. Acknowledgements

Author would like to thank to Vojtěch Svátek, Martin Labský and Martin Kavalec for their comments and help.

The research has been partially supported by the grant no. 201/03/1318 of the Grant Agency of the Czech Republic „Intelligent analysis of the WWW content and structure".

References

Berners-Lee T., Hendler J., Lassila O. (2001). *The Semantic Web*. *Scientific American*, vol. 284, no. 5, May 2001, p. 34-43

Brin, S. (1998). *Extracting Patterns and Relations from the World Wide Web*. In WebDB Workshop at EDBT'98

Brin S, Motwani R., Ullman J, Tsur S. (1997). Dynamic itemset counting and implication rules for market basket data. In SIGMOD 1997, *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255-264, Tucson, Arizona, USA, May 1997.

Kavalec M., Svátek V. (2002). Information Extraction and Ontology Learning Guided by Web Directory. In: *ECAI Workshop on NLP and ML for Ontology engineering (OLT-02)*. Lyon, 2002.

Maedche A., Neumann G., Staab S. (2002). Bootstrapping an Ontology Based Information Extraction System. Studies in Fuzziness and Soft Computing, editor Kacprzyk J., *Intelligent exploration of the web*, Springer 2002/01/01

Maedche A., (2002). Ontology Learning for the Semantic Web, Kluwer Academic Publ., ISBN: 0-7923-7656-0

Ratnaparkhi A.: *Adwait Ratnaparkhi's Research Interests*, [online], cited 2004
http://www.cis.upenn.edu/~adwait/statnlp.html

Riloff E., Jones R. (1999). Learning Dictionaries for Information Extraction by Mulit-Level Bootstrapping, *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, P 474-479, 1999

Salton G., Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 (5):513--523, 1988.

Uschold M., Jasper R. (1999). A Framework for Understanding and Classifying Ontology Applications. In: *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*.