Discovering Semantic Sibling Associations from Web Documents with XTREEM-SP

Marko Brunzel and Myra Spiliopoulou

Otto-von-Guericke-University Magdeburg
forename.name@iti.cs.uni-magdeburg.de

Abstract. The semi-automatic extraction of semantics for ontology enhancement or semantic-based information retrieval encompasses several open challenges. There are many findings on the identification of vertical relations among concepts, but much less on indirect, horizontal relations among concepts that share a common, a priori unknown parent, such as Co-Hyponyms and Co-Meronyms. We propose the method XTREEM-SP (Xhtml TREE Mining for Sibling Pairs) for the discovery of such binary "sibling"-relations between concepts of a given vocabulary. While conventional methods process an appropriately prepared corpus, XTREEM-SP operates upon an arbitrarily heterogeneous Web Document Collection on a given topic and returns sibling relations between concepts associated to it. XTREEM-SP is independent of domain and language and does not rely on linguistic preprocessing nor on background knowledge beyond the ontology it is asked to enhance. We present our evaluation results with two gold standard ontologies and show that XTREEM-SP performs well, while being computationally inexpensive.

1 Introduction

The discovery of semantic relations among terms is a crucial task in many applications on text retrieval and understanding. Ontologies, the backbone of the Semantic Web, rely on making semantic relations explicit. There are many methods for the discovery of vertical hierarchical relations. There is less work on the discovery of concepts that stand in a horizontal relation to each other and are the children of a common, not a priori known and possibly not interesting parent concept; “Co-Hyponym relations” and “Co-Meronym relations” are two types of such horizontal relationships. In this paper, we propose a method that identifies such *sibling relations*. In ontology engineering, there are different approaches for the discovery of semantic relations. Most of them [FN99, MS00, and BCM05] use unstructured plain text as input; semi-structured text is converted to plain text. There are also approaches that exploit resources like dictionaries, glossaries or database schemata [K99], but are limited to the rare case when such resources are available. Our method rather uses semi-structured content as input, exploiting the XHTML document structure.

The core of our method is XTREEM, a mechanism that performs Xhtml TREE Mining. In [BS06b], we have proposed XTREEM-SG that discovers groups of sibling concepts; an earlier version appeared in [BS06a]. In this paper, we extend the XTREEM core to find sibling pairs characterized by association strength, whereby the concepts come from a given vocabulary. XTREEM-SP does not use linguistic

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2006, LNCS 4081, pp. 469 – 480, 2006.

resources, nor a prepared corpus; it uses publicly available Web Documents. We show that XTREEM-SP finds pairs of concepts in Co-Hyponymy or Co-Meronymy relation with higher accuracy than conventional approaches.

In the next section, we discuss related work. In section 3, we present XTREEM-SP. Section 4 is devoted to experiments and evaluation using two gold standard ontologies from the domain of tourism. The last section concludes our study.

2 Related Work

The idea of using structural similarities [ZLC03, B04], including path structures, of XHTML/XML Documents is used for several goals, such as clustering documents on structural similarities [DCWS04, TG06, and CMK06]. In contrast we use the Path information to infer siblings. The constitution of the paths is not used itself; no comparison with paths from other documents is performed with XTREEM-SP.

The broad domain of research is *ontology learning*: A comprehensive overview on this subject has appeared recently in [BCM05]. Those approaches are focusing on ontology learning from text. There are also approaches performing *Ontology Learning from structure* [K99]: However, these methods use existing database schemas or other conceptualizations as input and are therefore limited to cases where such schemas are available, which is usually not the case. Closer related are studies also discovering semantics on the Web [FS02, AHM00].

Hearst patterns [H92] are used to find relations among terms in text collections. Also Co-Hyponym relations can be found with this approach. But the disadvantage is that such patterns are rare, the coverage is low, even on big document collections. Cimiano et al also discover (Co-)Hyponym relations by finding and analyzing examples of Hearst patterns on the WWW [CS04, CS05]. In [P05] instances of WordNet concepts are found within big Web Document Collections with a rule based mechanism ignoring the Mark-Up. The document structure is taken into account for the establishment of a knowledge base of extracted entities from the WWW in [ECD04].

The Acquisition of Co-Hyponym semantics from text with association measures is performed by [HLQ01], but there the document structure is not used. Kruschwitz [K01a, K01b] uses *Mark-Up* sections of Web Documents to learn a *domain model.* Similarly to our approach, Kruschwitz exploits the Mark-Up for the representation of similar concepts inside Web Documents. However, as opposed to our approach, the tree structure of (X)HTML documents is not incorporated. [ST04] uses also different tags of HTML documents for acquiring Hyponymy relations. They only use list *itemizations*. There is no mentioning of using the tree structure of (X)HTML documents in general, where contributions also from other tags than item elements can be expected.

3 Finding Sibling Groups with XTREEM-SP

XTREEM-SP is based on mark-up conventions that can be found in almost all Web Documents: Different authors use different nested tags to structure pieces of information in Web Documents, but tend to adhere to similar structures. XTREEM-SP exploits this observation to find terms appearing within the same syntactic structure of an XHTML (or HTML) document. Pairs of such terms are potentially correlated, so XTREEM-SP applies statistical to identify strongly associated pairs. Hence, XTREEM-SP can find pairs of correlated terms, even if they are not co-located inside the same narrow context window. This can be seen in the headings example of Table1: Both text spans "WordNet" and "Germanet" appear within the same syntactic structure, i.e. the sequence of HTML tags leading to them. Hence, XTREEM-SP uses such syntactic structures to infer semantic relatedness.

Table 1. Semantically related terms, located in different paragraphs or separated by other terms

Headings, located in different paragraphs	Highlighted keywords, separated by normal text

…
`<h2>WordNet</h2>`
`<p>Was developed …</p>`
`<h2>Germanet</h2>`
`<p>Analogous …</p>`
…
	`… <p> … there are different important standards for building the <strong>Semantic Web</strong>. … is <strong>RDF</strong>. … <strong>RDFS </strong> adds … whereas <strong>OWL </strong> is … </p> …`

The XTREEM-SP procedure, which aims to organize a given vocabulary of terms into Co-Hyponym groups, entails Pre-processing (Group-By-Path, the core of the XTREEM-SP approach) and Processing (Association Strength Calculation), which are shown in the following data–flow diagram (Fig. 1) and described in section 3.2.

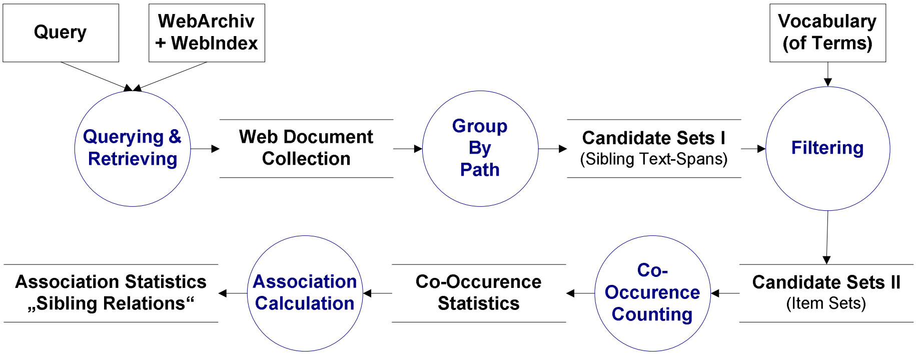

Fig. 1. Data-Flow Diagram of the XTREEM-SP procedure

We now introduce our algorithm XTREEM-SP that takes as input a collection of documents, observing each document as collection of Text-Span sets. On the elements of those sets a Co-Occurrence statistic is created. Upon this statistic association strength on term pairs is calculated. so that the terms with a strong association stand in sibling (Co-Hyponym, Co-Meronym) relationship to each other.

Step 1 – Querying & Retrieving: The XTREEM procedure operates on a *Web Document Collection*. Such a Web Document Collection is obtained by querying a *Archive+Index Facility* on a *query Q* with a Web Document Collection $W=\{d_1,...,d_s\}$ as result, for which Q is satisfied. Q constitutes the domain of interest whereupon semantics should be discovered. It should therefore encircle the Documents which are supposed to entail domain relevant content, e.g. "tourism*".

The Web Document Collection should be big enough to contain manifold occurrences of the desired concepts. The Web Document Collection is not supposed to be a small manually handcrafted document collection; bigger amounts of web content which have an appropriate coverage of the domain are more desirable. Here, recall is more important than precision. To obtain such a comprehensive Web Document Collection, alternatively a *focused web crawl* can be performed; when a vocabulary is given, this vocabulary can also be used to obtain Web Document references via the web services of internet search engines.

Step 2 - Group-By-Path: The Group-By-Path operation, described in detail in [BS06b] represents the core of the overall XTREEM-SP method. We consider Web Documents to find sibling relations among terms. We group Text-Spans that have the same Tag Path as its predecessor. The Group-By-Path approach performs a transition of a Web Document from a tree, to a collection of Pairs(Tag-Path, Text-Span) to a collection of Text-Span sets. For each $d_i \in W$ with $i=1,\ldots,s$ the Group-By-Path algorithm is applied. As result we obtain the collection of Text-Span sets $H'=(b_1,\ldots b_u)$.

Step 3 - Filtering: The aim of the procedure described in this publication is to infer semantically motivated sibling Pairs. Let $V=\{v_1,\ldots,v_p\}$ be the vocabulary of terms given as input. For the following steps we only consider all Text-Spans $e \in b$ which are contained in V. $H''=(b_1,\ldots,b_u)$ so that for all $e \in b$ it is also true $e \in V$.

Step 4 – Co-Occurrence Counting: In this step a Co-Occurrence statistic is created. Co-Occurrence is obtained from all pair wise occurrences of $e_1 \in b_i \cap e_2 \in b_i$ for all $b_i \in H''$ with $i=1,\ldots,u$. Such pairs are only obtained from b_i with cardinality > 1 since only sets containing at least two elements are able to reflect a sibling relation among their elements. For all Combinations of $v_1 \in V \cap v_2 \in V$ with $v_1 \neq v_2$ a count is associated reflecting how often a combination occurred in H''.

Step 5 – Association Calculation: From the counts on term pairs obtained in Step 4, the strength of the association between the pair components can be inferred in many ways. From simply using the raw Co-Occurrence frequency, through the many association measures from statistics (such as χ^2-Association [MS00]) to information theoretic measures (such as Mutual Information). For a comprehensive overview on association measures see [E04]. χ^2-Association is the association measure of our choice, since in the experiments it showed the best results and its application is appropriate on sufficiently large data sets as the ones obtained from big Web Document Collections.

4 Experiments and Evaluation

As evaluation reference we use two gold standard ontologies (GSO). The GSO's contain sibling relations besides other content. They also provide the closed vocabulary whereupon sibling relations are automatically derived by the XTREEM-SP procedure.

In Experiment 1 we will contrast the results obtained with XTREEM-SP (Group-By-Path) on sibling semantics against the results obtained on the traditional Bag-Of-Words vector space model and a further alternative method based on Mark-Up. In experiment 2 we will contrast the influence an Association Measure has compared to the solely usage of Co-Occurrence frequency. In experiment 3 we will investigate the influence of the input query which constitutes the Web Document Collection processed. In Experiment 4 we will vary the required minimum support of terms within the Web Document Collection to be processed.

4.1 Description of Experimental Influences

Evaluation Reference: The Evaluation is performed on two gold standard ontologies, from the tourism domain. The concepts of these ontologies are also terms, thus in the following the expressions “concepts” and “terms” are used interchangeably.

Sibling relations can be obtained from the GSO’s for all Sub-Concepts where the corresponding Super-Concept has more than one Sub-Concept; if there are at least two child Concepts of a Parent Concept. As a result, there is a number of Concept Pairs which stand in a sibling relation, whereas other Concept Pairs are not conceptualized as standing in sibling relation. We only use the direct Super-Concept Sub-Concept relation to derive sibling relations.

The “*Tourism GSO*”[1] contains 293 concepts grouped into 45 sibling sets resulting in 1176 concept pairs standing in sibling relation; the “*Getess annotation GSO*”[2] contains 693 concepts grouped into 90 sibling sets resulting in 4926 concept pairs standing in sibling relation.

There are three Inputs to the XTREEM-SP procedure described in the following:

Input(1) : Archive+Index Facility: We have performed a topic focused web crawl on “tourism” related documents. The overall size of the document collection is about 9.5 million Web Documents. The Web Documents have been converted to XHTML. With an n-gram based language recognizer non-English documents have been filtered out. The Documents are indexed, so that for a given query a Web Document Collection can be retrieved.

Input(2) : Queries: For our experiments we consider four document collections which result from querying the Archive+Index Facility. The constitution is given by all those documents adhering to Query1 - “touris*”, Query2 - “accommodation” and by the whole topic focused Web Document Collection reflected by Query3 – “*”. Additionally we give the results for Query4 – “accomodation”. Query4 was foremost misspelling on Query2, but since this variant is present in millions of Web Documents we will present theses results. Those variations are object of Experiment 3.

Input(3) : Vocabulary: The GSO’s described before, are lexical ontologies. Each concept is represented by a term. These terms constitute the vocabulary whereupon sibling relations are calculated.

The overall XTREEM-SP procedure is constituted of preprocessing and processing:

Procedure (1) : Preprocessing method: For the evaluation of the Group-By-Path sub procedure we will contrast our Group-By-Path (GBP) method with the traditional

[1] http://www.aifb.uni-karlsruhe.de/WBS/pci/TourismGoldStandard.isa

[2] http://www.aifb.uni-karlsruhe.de/WBS/pci/getess_tourism_annotation.daml

Bag-Of-Words (BOW) vector space model The BOW is the widespread established method on processing of textual data, while The variation of these influences is object Experiment 1.

Procedure (2) : Processing – Association Strength Derivation: From the raw sibling sets obtained by the Pre-processing, the Co-Occurrence frequency of Term Pairs is counted. This frequency can be used as indicator of association strength. We will refer to this method by “frequency”. With the χ^2-Association Measure, more statistical stable values of association strength can be calculated. The variation of these influences is object of Experiment 1 and Experiment 2.

In our experiments we found that some of the terms of the vocabulary are never or very rarely found on rather big Web Document Collections. E.g. one reference contains the errors “Kindergarden” instead of the correct English “Kindergarten”. To eliminate the influence of errors in the reference, we also vary the *required minimum feature support.* The support is given by the frequency of the features (terms) in the overall text of the Web Document Collection. We used minimum support thresholds from 0 (all features are used, nothing is pruned) to 100000 (0, 1, 10, 100, 1000, 10000, 100000). When the support is varied, only those features of the Vectorization and of the reference fulfilling these criteria are incorporated into the evaluation. The variation of these influences is object Experiment 4.

4.2 Evaluation Criteria

From the gold standard ontologies we extract all Concept Pairs which stand in a sibling relation to each other. This is in the following also referred to as “Reference”.

Object of the evaluation is a ranked list of automatic obtained Concepts Pairs, whereas the ranking is given according to the Association Strength of the Concept Pair. For each automatic obtained Concept Pair can be determined if this relation is also supported by the Reference which gives a positive count. If a Concept Pair is not supported be the Reference a negative count is assumed. With this, for each position in the ranked list, recall and precision can be calculated. The recall is the number of already seen true Sibling Pairs (#positive) to the number of Sibling Pairs given by the Reference (#overall). The precision is the number of true Sibling Pairs (#positive) to the number of seen automatic generated Pairs (#positive + #negative).

$$recall = \frac{\#\,positive}{\#\,overall} \qquad precision = \frac{\#\,positive}{\#\,positive \;+\; \#\,negative}$$

For a ranked list of associated Term Pairs a recall precision chart line can be obtained by a series of measurements on recall precision values.

4.3 Experiments

In the following we will show the results obtained from the experiments. Table 2 shows the number of documents which adhere to a certain query. This corresponds to the size of the Web Document Collection which is processed by the subsequent following processing steps. Table 2 also shows the number of candidate sibling sets obtained after performing the Pre-processing on different Queries for the two vocabularies. Only terms which are present in the input vocabulary are observed in the subsequent. Table 2 also shows the number of observed pairs derived from these sets.

Table 2. Experimental Data Numbers

Query Name	Query Phrase	Number of Documents	Number of Candidate Sets II obtained with GBP		Number of Sibling Pairs (from Candidate Sets II)	
			GSO1	GSO2	GSO1	GSO2
Query1	"touris*"	1,468,279	222,037	318,009	1,600,440	3,804,214
Query2	"accommodation"	1,612,108	293,225	373,802	2,092,432	3,885,532
Query3	"*"	9,437,703	924,045	1,326,843	5,763,596	14,071,016
Query4	"accomodation"	471,540	78,289	98,886	686,108	1,198,224

Experiment 1: Group-By-Path in comparison to alternatives Methods

In this experiment we will contrast the quality of results on sibling relations obtained with the Bag-Of-Words (BOW) vector space model, on a usage of Mark-Up without Path Information as described in [K01a] against our new Group-By-Path method. Query1 was chosen as the query constituting the Web Document Collection. The comparison was performed for two methods on association strength (frequency, χ^2) and for both references (GSO1,GSO2).

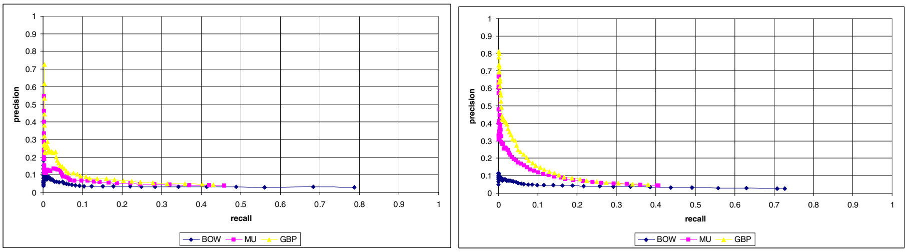

Fig. 2. and 3. Pre-processing - BOW vs. MU vs. GBP (Frequency,Query1) for GSO1 and GSO2

The diagrams which result on the usage of "frequency", Fig. 2 and Fig. 3, show that GBP performs best for both GSO's. MU performs better than BOW. The overall measured results are relatively low. On the top ranked association Pairs, GBP (and MU) shields a high precision which then rapidly declines. For higher recall values the chart lines converge. Since a recall above 40 percent is only obtained on BOW, we can conclude that some sibling relations never occur up on Marked-Up Web Document Structure. This does not necessarily mean that GBP is weak; since the ontologies do not directly encode sibling relations, there may exist Concepts which tend not to occur together. E.G. "ski school" and "surf school" may be sub-concepts of "sport school" but are rather unlikely to be discovered from content. The evaluation criteria can not prevent from such cases.

Fig. 4 and Fig. 5 show the results by using the association strength calculated by the χ^2-Association Measure. In contrast to the usage of "frequency", the results of MU are nearly the same as for GBP. An explanation for this is that the χ^2-Association Measure here performs well on diminishing sporadic occurrences which can happen on MU in comparison to GBP. BOW performs again worst. All the experiments

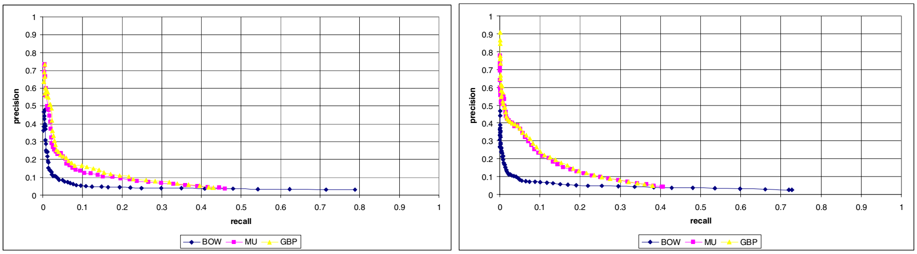

Fig. 4. and 5. Pre-processing - BOW vs. MU vs. GBP (χ^2,Query1) for GSO1 and GSO2

within this publication are performed on a closed vocabulary. The choice of Pairs observed in the documents is therefore drastically limited in comparison when using an open vocabulary. When using an open vocabulary the alignment of association generated with GBP towards sibling semantics, in comparison to MU, becomes more visible than measured on the limit vocabulary.

Conclusion: Our experiments on automatically obtaining sibling relations showed that our Group-by-Path method, the core of the XTREEM-SP procedure, shows the best results. Though it was not claimed that the Bag-Of-Words model is strong on capturing sibling semantics, we can confirm our hypothesis that the results obtained with XTREEM-SP (based on GBP) are motivated by sibling semantics.

Experiment 2: Different Methods on Association Strength in Comparison

In this Experiment we will focus on how variations on the method association strength is obtained influences the results. Specifically we will use the Co-Occurrence frequency and the χ^2-Association Measure [MS00]. In Experiment 1 for the different association strength methods this was done in series; in contrast Fig. 6 shows the chart lines on GBP of Fig. 2 and Fig. 4 together. Fig. 7 shows the chart lines on GBP of Fig. 3 and Fig. 5 together.

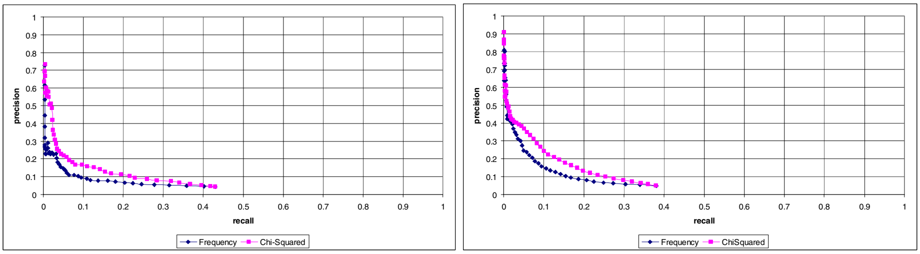

Fig. 6. and 7. Association Strength - Frequency vs. χ^2 (GBP,Query1) for GSO1 and GSO2

Fig. 6 and Fig. 7 show that on both vocabularies/references the usage of χ^2 - Association strength shielded the best results.

We also used Mutual Information and Poison Sterling Association Measure as well as cosines distance; the results are comparable to χ^2-Association or worse but better than just frequency. The literature on the quality of these association measures mentions that different association measures perform sometimes better, sometimes worse than other with no clear conclusions. In the experiments of this publication χ^2-Association Measure gave the best results compared to the solely frequency support.

Experiment 3: Varying the Topic Focus

XTREEM-SP relies on constituting a Web Document Collection by a query. A query therefore represents the focus of the data analyzed. Here we will investigate how variations on the query influence the obtained results on sibling semantics. The different Queries are shown in Table 2.

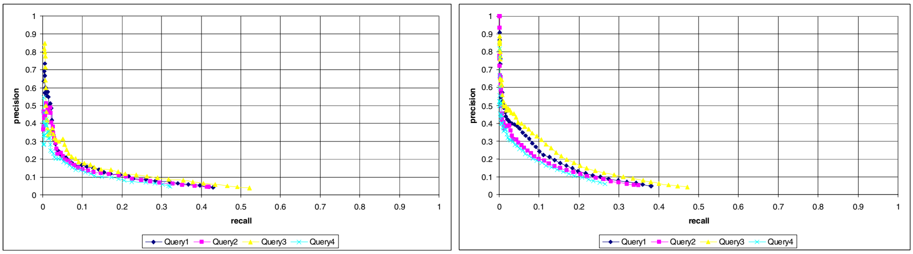

Fig. 8. and 9. Results on different Web Document Collection constituting Queries (GBP, χ^2) for GSO1 and GSO2

As Fig. 8 shows, the results of all for Queries are closely together for GSO1. For GSO2 the results vary more than for GSO1. For both GSO's, Query3 – "*" which depicts the full topic focused Web Document Crawl shielded the best results. A explanation for this is that with the single phrase queries (Query1,Query2 and Query4) always a too focused Web Document Collection is processed. The Reference contains terms – and relations which are not present on Web Documents adhering to a certain "focused" query. This means that for practical settings a combined query (E.G. "touris* OR accommodation OR holidays OR 'sport event' … ") may be the better choice. On the other hand a ontology engineer will rather focus on a fraction of the conceptualization to be obtained or improved at one moment and therefore focused Queries are appropriate.

Experiment 4: Variations on the required support

In the last experiment we will investigate the influence of the term frequency in the Web Document Collection on the obtained results. As a side effect of an increased required support, "misconceptualization", present in the reference ontologies, is outweighed. With increasing required support more and more relations are not relevant, which is reflected by eliminating these Pairs from the reference. Table 3 shows the decreasing number of relations by increased required term support. We used the support of terms, not of the Co-Occurrence of term Pairs which would be an alternative approach. As Fig. 10 and Fig. 11 show, for increased required support, better results regarding recall and precision are obtained. This means that recall and precision on sibling relations of high frequent terms are found better than on low frequent ones.

Table 3. Decreasing number of reference sibling relations on increased required support

Required support		0	1	10	100	1000	10000	100000
Number of reference sibling relations	GSO1	1176	1120	1033	844	637	404	161
	GSO2	4926	4553	4073	3439	2653	1006	582

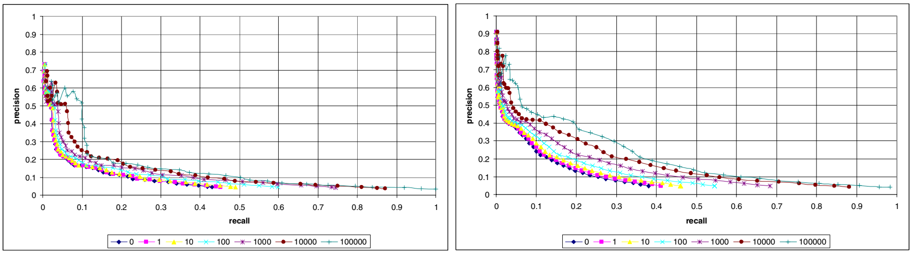

Fig. 10. and 11. Variations on the Required Support (Query1,GBP, χ^2) for GSO1 and GSO 2

5 Conclusions and Future Work

We have presented XTREEM-SP, a method that discovers binary horizontal semantic relations among concepts by exploiting the structural conventions of Web Documents XTREEM-SP processes Web Documents collected from the WWW and thus eliminates the need for a well-prepared document corpus. Furthermore, it does not rely on linguistic pre-processing or NLP resources. So, XTREEM-SP is much less demanding of human resources. Our experiments with two golden standard ontologies and with several parameter variations show that XTREEM-SP delivers good results, i.e. semantically meaningful sibling pairs.

Our method is only a first step on the exploitation of the structural conventions in Web Documents for the discovery of semantic relations. In our future work we want to investigate the impact of individual Mark-Up element tags like <p>, <li>, and <dt> on the results. Discovering the corresponding Super-Concept for the Sub-Concepts standing in sibling relation is a further desirable extension.

References

[AHM00] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the WWW, Proc. of the Workshop on Ontology Construction ECAI-2000

[B04] D. Buttler. A short survey of document structure similarity algorithms. In Proc. of the *International Conference on Internet Computing*, June 2004.

[BCM05] P. Buitelaar, P. Cimiano, Bernardo Magnini, Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence and Applications Series Volume 123, IOS Press, Amsterdam, 2005

[BS06a] M. Brunzel, M. Spiliopoulou. Discovering Multi Terms and Co-Hyponymy from XHTML Documents with XTREEM. In Proc. of *PAKDD Workshop on Knowledge Discovery from XML Documents (KDXD 2006)*, LNCS 3915, Singapore, April 2006

[BS06b] M. Brunzel, M. Spiliopoulou. Discovering Semantic Sibling Groups from Web Documents with XTREEM-SG. In Proc. of *EKAW* 2006 (accepted for publication), Podebrady, Czech Republic, October 2006

[CMK06] I. Choi, B. Moon, H-J- Kim. A Clustering Method based on Path Similarities of XML Data. Data & Knowledge Engineering, vol. no. pp.0-0, Feb. 2006

[CS04] P. Cimiano and S. Staab. Learning by googling. *SIGKDD Explorations*, 6(2):24-34, December 2004.

[CS05] P. Cimiano, S. Staab. Learning concept hierarchies from text with a guided hierarchical clustering algorithm. *Workshop on Learning and Extending Lexical Ontologies at ICML 2005*, Bonn 2005.

[DCWS04] T. Dalamagas, T. Cheng, K. J. Winkel, T. Sellis, Clustering XML documents using structural summaries, in Proc. of the *EDBT Workshop on Clustering Information over the Web (ClustWeb04)*, Heraklion, Greece, 2004

[E04] Stefan Evert, The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, University of Stuttgart. 2004

[ECD04] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Web-Scale Information Extraction in KnowItAll. Proc. of the *13th International WWW Conference*, New York, 2004

[FN99] D. Faure, C. Nedellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: the system ASIUM, In Proc. of *EKAW 1999*

[FS02] A. Faatz, R. Steinmetz, Ontology Enrichment with Texts from the WWW, Proc. of the *First International Workshop on Semantic Web Mining, ECML 2002*, Helsinki 2002

[H92] M. Hearst, Automatic acquisition of hyponyms from large text corpora. In Proc. of the *14th International Conference on Computational Linguistics*, 1992

[HLQ01] G.Heyer; M. Läuter, U. Quasthoff, Th. Wittig, Ch. Wolff. Learning Relations using Collocations. In Proc. *IJCAI Workshop on Ontology Learning, Seattle/WA*, 2001

[K01a] U. Kruschwitz, A Rapidly Acquired Domain Model Derived from Mark-Up Structure. In Proc. of the *ESSLLI'01 Workshop on Semantic Knowledge Acquisition and Categorization*, Helsinki, 2001.

[K01b] U. Kruschwitz. Exploiting Structure for Intelligent Web Search. Proc. of the *34th Hawaii International Conference on System Sciences (HICSS)*, Maui Hawaii 2001, IEEE

[K99] V. Kashyap. Design and creation of ontologies for environmental information retrieval. Proc. of the *12th Workshop on Knowledge Acquisition, Modeling and Management*. Alberta, Canada. 1999.

[MS99] C. Manning, H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA: May 1999.

[MS00] A. Maedche and S. Staab. Discovering conceptual relations from text. In Proc. of *ECAI 2000*

[P05] M. Pasca. Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded. In: *CICLing-2005*, LNCS 3406, 2005.

[ST04] K. Shinzato and K. Torisawa. Acquiring hyponymy relations from Web Documents. In Proc. of the 2004 *Human Language Technology Conference (HLT-NAACL-04)*, Boston, Massachusetts, 2004.

[TG06] A. Tagarelli, S. Greco. Toward Semantic XML Clustering. *6th SIAM International Conference on Data Mining (SDM '06)*. Bethesda, Maryland, USA, April 20-22, 2006

[ZLC03] Z. Zhang, R. Li, S. Cao, and Y. Zhu. Similarity metric for XML documents. In Proc. of the *Workshop on Knowledge and Experience Management*, October 2003.