# Phrase Pair Classification for Identifying Subtopics

Sujatha Das[1], Prasenjit Mitra[2], and C. Lee Giles[2]

[1] Department of Computer Science and Engineering
[2] School of Information Science and Technology
The Pennsylvania State University
University Park, PA-16802
{gsdas@cse,pmitra@ist,giles@ist}.psu.edu

**Abstract.** Automatic identification of subtopics for a given topic is desirable because it eliminates the need for manual construction of domain-specific topic hierarchies. In this paper, we design features based on corpus statistics to design a classifier for identifying the (subtopic, topic) links between phrase pairs. We combine these features along with the commonly-used syntactic patterns to classify phrase pairs from datasets in Computer Science and WordNet. In addition, we show a novel application of our *is-a-subtopic-of* classifier for query expansion in Expert Search and compare it with pseudo-relevance feedback.

**Keywords:** hypernym classification, expert search, query expansion.

## 1 Introduction

Domain-specific topic/concept hierarchies are used in several Information Retrieval (IR) related tasks such as document classification, clustering and visualization. Automatic techniques for inducing concept hierarchies that avoid manual effort from domain experts were studied before [10,6,12]. In this paper, we address the problem of identifying (subtopic, topic) phrase pairs in a particular domain. For instance, "support vector machines" *is-a-subtopic-of* "machine learning" and "query optimization" *is-a-subtopic-of* "query processing". This problem, is very similar to the hypernym classification from Information Extraction which involves extracting (instance, class) pairs such as $(apple, fruit)$ or $(France, country)$. In this paper, we use a representative corpus from the relevant domain to design features that indicate *is-a(-subtopic-of)* links among phrase pairs. We illustrate the results of our classifier on two datasets: one from the Computer Science domain and the second from WordNet. In addition, we demonstrate preliminary query expansion results for Expert Search using our subtopic identifier. Expert search involves ranking people based on their expertise on a particular (queried) topic. Based on the intuition that expertise on subtopics is indicative of expertise on the general topic, our classifier extracts expansion subtopic phrases in the pseudo-relevance loop to obtain better retrieval performance.

R. Baeza-Yates et al. (Eds.): ECIR 2012, LNCS 7224, pp. 489–493, 2012.

We briefly summarize previous related research in Section 2. Design of our features and query expansion for expert search is discussed in Section 3 whereas experimental datasets and results are presented in Section 4.

## 2 Related Work

Generating concept hierarchies using document subsumption was studied by Sanderson, et al [10]. Terms extracted using these techniques were used to enhance the end-user browsing experience [6,12]. Begelman studied term clustering techniques for tagging [1,7]. Hypernym classification or is-a mining is a well-studied information extraction problem mostly solved with the use of lexical and syntactic patterns [4,3,11]. Our application task, Expert Search, was extensively studied in TREC[1]. We use the probabilistic models extended for expert finding in scientific collections [2] as our baseline and illustrate query expansion in this framework. Macdonald et al. studied query expansion for expert search using techniques involving pseudo-relevance feedback [9,8].

## 3 Subtopic Identification

Given a corpus (document collection) representative of a domain and a couple of phrases, $t$ and $p$ our goal is to design features to classify the pair $(p, t)$ as positive or negative depending on whether $p$ *is-a-subtopic-of* $t$. For example (routing algorithms, computer networks) is a positive pair whereas (routing algoritms, natural language processing) is not. Let $D_t$ and $D_p$ be the sets of documents in which $t$ and $p$ respectively occur in the corpus. If $t$ and $p$ are related we can expect to see an overlap in the sets of documents they occur in. Similarly, we can expect an overlap in the terms used in the document sets, $D_t$ and $D_p$ since $p$ and $t$ are likely to occur in similar contexts. The PMI-IR (Pointwise Mutual Information in Information Retrieval) measure proposed by Turney and given by $\frac{|D_t \cap D_p|}{|D_t| * |D_p|}$ corresponds to the statistical dependence between $t$ and $p$ in terms of their occurrence. Similarly, if $t$ is more general than $p$, we can expect $IDF(p) > IDF(t)$ where IDF or inverse document frequency is a measure of a term's rareness in a corpus. Note that this feature also checks if $|D_t| > |D_p|$ since $IDF(w) = log(1 + \frac{N_c}{N_w})$ where $N_c$ is the number of documents in the entire collection and $N_w$ is the number of documents containing the word $w$. This variant is easier to satisfy in incomplete collections than Sanderson et al.'s $D_p \subseteq D_t$ [10]. In addition, we can also expect the average similarity of documents in $D_t$ to be less than average similarity of documents in $D_p$ since $t$ is more general. Lexico-syntactic patterns are widely used to to form queries with class names on the Web, in the is-a mining task. For instance, if a sentence fragment containing "cities like Paris" is seen several times, one can reasonably conclude that *Paris is-a city*. We augment our corpus-based features with these syntactic features for our classification task (Table 1). NumWebHits(r) in the table refers to the

[1] http://trec.nist.gov

**Table 1.** Features for classifying the pair $(p, t)$

| | Feature Description |
|---|---|
| 1-3 | IDF(p), IDF(t), IDF(p)/IDF(t) |
| 4-6 | $AvgSim(D_p)$, $AvgSim(D_t)$, $AvgSim(D_p)/AvgSim(D_t)$ |
| 7-8 | $Sim(First(D_p), First(D_t))$, $Sim(Last(D_p), Last(D_t))$ |
| 9,10 | $PMI - IR(p, t)$, $Sim(D_p, D_t)$ |
| 11-16 | Are NumWebHits with pattern r $> 0$ for $r \in R$? |
| 20-22 | Are NumWebHits(r) $> 0$ for at least one/two/three $r \in R$? |

number of hits on the web using the pair $(p, t)$ along with a syntactic pattern $r$, where $r \in R = \{$"such as", "including", "like", "and other", "or other" and "especially" $\}$. For instance, "cities `including` Paris" forms an example web query. $First(D_p)$ and $Last(D_p)$ refer to the first and the last documents in $D_p$ witn documents ranked based on their retrieval scores using $p$ as the query.

## 4 Experiments and Results

We evaluated our *is-a(-subtopic-of)* classifier on a data set of Computer Science topic pairs collected from EventSeet[2]. In addition, we also evaluated the classifier performance for identifying hypernyms in the open-domain using pairs from WordNet[3]. We excluse the details of these dataset creation due to space limitations but will be made publicly available (after the blind review). For the EventSeer dataset, the CiteSeerX[4] document collection was used as a corpus whereas for the WordNet dataset, Wikipedia abstracts form the representative corpus[5]. Both the datasets contain around 400 positive and 400 negative pairs.

We used Indri[6] and Yahoo BOSS[7] for extracting features whereas the classifier was trained using $SVM^{light}$[8]. The average five-fold classificiation accuracy was 0.72 on the EventSeet dataset and 0.68 on the WordNet dataset. Using corpus-based features alone, the accuracies were 0.53 on the EventSeer set and 0.64 on the WordNet set.

### 4.1 Example Application: Query Expansion for Expert Search

An Expert Search system ranks authors based on their expertise on the queried topic. For a scientific document collection, authorship of "relevant" documents on a topic is treated as evidence of expertise on that topic. Consider a user searching for experts on "artificial intelligence". From common knowledge, a

[2] http://eventseer.net

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

[4] http://citeseerx.ist.psu.edu

[5] http://en.wikipedia.org/wiki/Wikipedia_database

[6] http://www.lemurproject.org/

[7] http://developer.yahoo.com/search/boss/

[8] http://svmlight.joachims.org/

| Original Query | Set-1:our subtopic phrases, Set-2:stemmed expansion terms from Indri |
|---|---|
| Information Extraction | { wrapper induction, tree automata, seed tuples, extraction rules, scalable information extraction, named entity}<br>{ extract, inform, system, text, document, process, web, pattern, gener} |
| Support Vector Machines | { combination of kernels, procrustes kernel, kernel functions, statistical learning, neural information, feature selection}<br>{ vector, machin, support, neural, classif, method, network, featur} |
| Intelligent Agents | {agent technology, supply chain, agent communication, mobile agents}<br>{agent, intellig, system, network, manag, inform, commun, environ, user, distribut} |

person's expertise on subtopics of "artificial intelligence" such as "machine learning" and "robotics" indicates some expertise in the main topic, "artificial intelligence". Based on this intuition, we tested the effect of using subtopic phrases for query expansion in Expert Search. The classifier was folded into the pseudo-relevance feedback loop and evaluated on the expert search queries from the ArnetMiner dataset[9] and the CiteSeerX collection. Deng's probabilistic model was used for ranking experts on the documents retrieved with Indri [2]. The baseline model obtained a precision and recall of 0.1286 and 0.0406 on this set. Using pseudo-relevance feedback as proposed by Lavrenko [5] improved these numbers to 0.1429 and 0.0454 with 50 expansion terms. Using number of terms less than 50 (i.e. $10, 20, 30$) did not yield improvements in terms of aggregate precision and recall. In contrast, adding just the top-5 subtopics identified by our method resulted in a precision of 0.1857 and a recall of 0.0586. Although this evaluation was on a small dataset (seven queries), anecdotal examples of expansion phrases shown in the table below indicate the potential of our classifier in identifying relevant subtopic phrases.

## 5 Conclusions and Future Work

We showed that a representative document collection from a domain can be used to identify (subtopic, topic) phrase pairs in that domain. Our preliminary results indicate that query expansion using subtopic phrases is more effective for expert search than directly employing pseudo-relevance feedback. For future work, we seek to improve the accuracy of our classifier by identifying other features that capture the (subtopic, topic) connection with higher precision. In addition, the relative effectiveness of each feature needs to be studied using feature selection techniques. Our final goal is to use an accurate *is-a-subtopic-of* classifier as a building block in extracting topic taxonomies and for effective query expansion.

## References

1. Begelman, G., Keller, P., Smadja, F.: Automated tag clustering: Improving search and exploration in the tag space. In: WWW (2006)
2. Deng, H., King, I., Lyu, M.R.: Formal models for expert finding on dblp bibliography data. In: ICDM (2008)
3. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in knowitall: (preliminary results). In: WWW (2004)

[9] http://arnetminer.org/

4. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING (1992)
5. Lavrenko, V., Croft, W.B.: Relevance based language models. In: SIGIR (2001)
6. Lawrie, D., Croft, W.B.: Discovering and comparing topic hierarchies. In: RIAO (2000)
7. Lin, H., Davis, J., Zhou, Y.: An Integrated Approach to Extracting Ontological Structures from Folksonomies. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 654–668. Springer, Heidelberg (2009)
8. Macdonald, C., Ounis, I.: Expertise drift and query expansion in expert search. In: CIKM (2007)
9. Macdonald, C., Ounis, I.: Using Relevance Feedback in Expert Search. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 431–443. Springer, Heidelberg (2007)
10. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: SIGIR (1999)
11. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: NIPS (2005)
12. Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S.: Discovering subsumption hierarchies of ontology concepts from text corpora. Web Intelligence (2007)