Web People Search with Domain Ranking

Zornitsa Kozareva[1], Rumen Moraliyski[2], and Gaël Dias[2]

[1] University of Alicante, Campus de San Vicente, Spain
zkozareva@dlsi.ua.es
[2] University of Beira Interior, Covilhã 6201-001, Portugal
rumen@hultig.di.ubi.pt, ddg@di.ubi.pt

Abstract. The world wide web is the biggest information source which people consult daily for facts and events. Studies demonstrate that 30% of the searches relate to proper names such as organizations, actors, singers, books or movie titles. However, a serious problem is posed by the high level of ambiguity where one and the same name can be shared by different individuals or even across different proper name categories. In order to provide faster and more relevant access to the requested information, current research focuses on the clustering of web pages related to the same individual. In this paper, we focus on the resolution of the web people search problem through the integration of domain information.

1 Introduction and Related Work

Most current web searches relate to proper names. However, there is a high level of ambiguity where multiple individuals share the same name and thus the harvesting and the retrieval of relevant information becomes more difficult. This issue directed current researchers towards the definition of a new task called Web People Search (WePS) [7]. The main idea of the task is to be able to group together e.g., cluster web pages that relate to different people sharing the same name.

Prior to the WePS task [2], [13] focused on similar task called name discrimination. Name discrimination stands for being able to identify the different individuals behind a name, but since the number of clusters is unknown, they called the task discrimination. [13] and [9] focused on the discrimination of cities, countries, professors among other categories and conducted evaluation in Spanish, Romanian, English and Bulgarian languages. The obtained results are promising showing that the discrimination can be performed not only on multi categories but also on multilingual level. The approach of [13] used co-occurrence vectors that represent the word co-occurrences derived from the document with the ambiguous name. The hypothesis is that names appearing in different documents and having similar co-occurrence context vectors, are highly probable to refer to the same individual. [12] reported the adaptation of the vector co-occurrence approach in a web environment. [8] applied semantic information derived from latent semantic indexing and relevant domains in order to estimate the similarity among snippets containing ambiguous names.

In the WePS challenge, there were 16 participants all using different sources of information. The CU-COMSEM [4] system ranked first and it was using token-based and phrase-based information. The modeled features represent the importance of the word in a document and in the whole document collection, the urls, the overlapping of noun phrases and named entities (NE). In contrast to this rich feature based system, the second best performing system IRST-BP [14] used only information from the NEs and their co-reference chaining. UBC-AS [1] system used word co-occurrence information to build a graph. The web page clustering was based on the HITS ranking. The system ranked on final position which indicates that simple word co-occurrence information is not sufficient for the WePS task.

Our main motivations in this paper are to develop a graph-based ranking criteria for synsets in a domain. The approach is flexible to the changing of the number of domain labels or the addition and deletion of synsets in *WordNet*. The second objective is to use the domain information for the resolution of WePS where each word is represented by its corresponding domains and associated weights. Web pages with similar domains are grouped together, because our hypothesis is that these pages are most likely to refer to the same individual.

The rest of the paper is organized as follows: in Section 2 we describe the *WordNet Domains* and the *eXtended WordNet* resources, Section 3 focuses on the graph construction and the synset ranking. Section 4 describes the integration of the domain information for the resolution of WePS. The approach is evaluated in Section 5 and we conclude in Section 6.

2 Domain Information

2.1 WordNet Domains

WordNet Domains[3] is a lexical resource developed by Magnini et. al, [11]. There are 167 hierarchically organized domain labels which are assigned to the *WordNet 2.0* synsets in semi-automatic way. For instance, the noun *bank* with senses 1 and 3 denoted as *bank#n#1* and *bank#n#3* belongs to the domain `ECONOMY`, *bank#n#9* belongs to the domain `ARCHITECTURE`, while *bank#n#2* belongs to the domains `GEOGRAPHY` and `GEOLOGY`. An important property of *WordNet Domains* is the ability to group together under the same domain words from different syntactic categories such as the verb *operate#v#7* and the nouns *doctor#n#1* and *hospital#n#1* under the domain `MEDICINE`. The domains act also as semantic bridges between words which might be unrelated or far from each other in the *WordNet* hierarchy. Unfortunately, current *WordNet Domains* does not provide information about the importance of a domain to a word. Therefore, we decided to introduce a graph-based criteria which ranks a set of concepts according to their importance within a domain. Our main motivation for graph ranking is the flexibility of the technique in terms of future changes of the number of domains or the addition/deletion of synsets.

[3] http://wndomains.itc.it/wordnetdomains.html

2.2 eXtended WordNet

For the creation of the graph, we used the information of the synset and the words in the gloss of the synset. According to [6], if a synset s_i has positive (negative) meaning, it can be interpreted as an indication that the synsets s_k defined by the terms occurring in the gloss of s_i are themselves with positive (negative) meaning. We adapted the idea for the domain ranking, assuming that the domains of the words in the gloss determine the domain of the synset and vice verse.

To create the graph, we used *eXtended WordNet*[4] because the words in the gloss are disambiguated and lemmatized. *eXtended WordNet* has been automatically generated, which means that the associations between the words and the synsets are likely to be incorrect sometimes and this can bring noise to our graph-based ranking method.

3 Capturing the Importance of a Synset in a Domain

3.1 Synset Graph Representation

We construct a graph G=$< N, E >$, with vertex N corresponding to the *eXtended WordNet* synsets s_i and the synsets from the gloss definitions s_k. A directed edge E represents the link from s_i to s_k if the gloss of s_i contains a term in s_k. The graph is constructed from nouns, verbs, adjectives and adverbs, because they carry out the domain information. The graph has no outcoming edges for vertex whose synsets have missing glosses.

3.2 Synset Ranking with PageRank

In order to determine which synsets s_i are more important for a domain D_p, we use the well known PageRank (PR) algorithm [3]. PageRank is a random-walk model initially developed for Web page ranking. It assigns numerical weighting to the vertices in the graph with the purpose of "measuring" the relative importance within the set.

The main idea of PR is like "voting". When one vertex links to another vertex, it casts a vote. The higher the number of votes a vertex has, the higher its importance becomes. Moreover, the importance of a vertex casting a vote determines how important the vote itself is. The score of PR associated with a node N is defined as:

$$PR(N) = (1-d) * e_j + d * \sum_{N_i \in E} \frac{PR(N_i)}{outdegree(N_i)} \quad (1)$$

where $d \in [0;1]$ is the so called damping factor and it is usually set to 0.85. The $outdegree(N_i)$ corresponds to the number of outgoing edges from vertex N_i, E

[4] xwn.hlt.utdallas.edu/

is the set of vertices having edge directed to N. The values e_j sum up to unity. We set this variables to be non zero for the senses of the domain for which PR is calculated and 0 for the rest of the words senses. Thus the words from the domain have greater impact on the ranking that the others.

The algorithm is iterative, initially assuming that each node has equal weight and throughout the iterations, the vertices accumulate weight. When the algorithm converges, e.g. the weights of the vertices between two consecutive iterations change less than a certain threshold, PR lists in descending order the important synset $s_1, .., s_k$ for domain D_p. We ran the algorithm for each one of the *WordNet Domains* setting e_j.

For instance, for the domain `ECONOMY` synset s_{10} occurs in the glosses of the synsets s_1, s_4 and s_5. As s_1, s_4, s_5 belong to the domain `ECONOMY` and have big weights, we assume that it is highly probable for s_{10} to belongs to the domain `ECONOMY`. Finally, for each domain we have a list of the synsets and a PR score which indicates how probable for the synset is to belong to the domain. The PR domain approach can be used as word sense disambiguation or sense reduction technique. For instance, the word *bank#n* has ten senses and seven domains, but PR estimated that *bank* is significantly important to the domain `ECONOMY` and but not so informative in the other six domains.

4 WePS with Domain Information

For each ambiguous person name, we have 100 web pages. The first step of our WePS approach consists in removing the *html* tags. Then the pages are tokenized, split into sentences and the words are POS tagged using the Montylingua toolkit [10]. We have automatically discarded sentences with length of one or two words, because they tend to refer to web links or menu titles.

4.1 WePS Domain Ranking

For each document in the ambiguous person name collection, we map the words to their corresponding domains. In order to determine the most relevant domains for a document, we introduce three different domain ranking schemes:

tf*idf: each sense s_{p,w_i} of word w_i in the documents is substituted by its *WordNet Domains*. For each domain we calculate tf*idf score which measures the relevance of the domain in the document. Evidently, a domain that is seen in all documents has low tf*idf score and is considered uninformative for the document and the whole document collection.

binary vector: each sense s_{p,w_i} of word w_i in the documents is substituted by its *WordNet Domains*. The final score of a domain in a document is the cumulative count of all words that pointed to the domain.

PR: for each sense s_{p,w_i} of word w_i in the documents, we consider only the domain with the highest PR weight. The rest of domains are discarded.

4.2 Document Similarity

In order to estimate the proximity of the web pages, we used the *cosine* and *Euclidean* distance measures. Our hypothesis is that web pages with similar domains and highly probable to refer to the same individual.

Cosine is a measure of similarity between two vectors of n dimensions by finding the angle between them. The number of dimensions in our case corresponds to the number of domains in the document collection. Given two web page vectors $WP(A)$ and $WP(B)$, the cosine similarity is the dot product $cosine(WP(A), WP(B)) = \frac{WP(A) \bullet WP(B)}{|WP(A)||WP(B)|}$. The value of 1 indicates that the two documents are very similar, while 0 indicates the documents are dissimilar.

The Euclidean distance between two web pages is the length of the path connecting them: $d = |WP(A) - WP(B)| = \sqrt{\sum_{i=1}^{n} |WP(A)_i - WP(B)_i|^2}$. In order to convert the Euclidean distance to similarity measure, we normalized it using the biggest distance and subtracted by one.

Both measures are calculated for all pairs of documents. From the obtained values we construct $N \times N$ dimensional matrix, where N corresponds to the number of document for an ambiguous name. In the WePS challenge, N is 100.

4.3 Document Clustering

The final step of our WePS algorithm consists in the clustering of the web pages according to their representation in the domain space. For the purpose we used PoBOC clustering [5] which builds a weighted graph with weights being the distances among the objects. Afterwards tight components, called poles, are found and the rest of the elements are assigned to the set of most similar poles. Thus, PoBOC decides on the number of clusters without need of a predefined threshold. PoBOC is a soft clustering algorithm and its application for the WePS tasks seems natural.

5 Experimental Evaluation

5.1 Data Description

The WePS task emerged as part of the Semeval 2007 challenge. The test data set[5] consists of 30 ambiguous person names. For each name, the organizers collected the first 100 web pages from Yahoo search engine or Wikipedia. The collected web pages have been manually annotated denoting whether a web page corresponds to given person or not. The number of clusters corresponds to the number of senses a person name has. In order to evaluate the performance of our system, we have used the official WePS evaluation script which measures the purity, inverse purity and f-score of the system.

[5] http://nlp.uned.es/weps/task-data.html

5.2 Experimental Results

Table 1 shows the obtained WePS results for the test data set. Our systems are denoted with the initials of the different domain ranking schemes and the used document similarity measure. In the table, we show the two best and the last ranked systems in the WePS challenge as well as two rankings provided by the organizers. The ONE-IN-ONE stands for one document per cluster and has the lowest possible inverse purity for this particular data set. While the ALL-IN-ONE means that all web pages are taken to be related to the same person i.e. only one cluster, and gives the lowest possible purity measure. Observing the performance of the ONE-IN-ONE baseline, the WePS participants discussed that it is not a satisfactory result to show to the final user single pages, because even systems below the ONE-IN-ONE baseline clustered the web pages more appropriately and informatively. In the future more sensitive evaluation measures are needed.

System	**Purity**	**Inverse Purity**	**F** $\alpha = 0.5$
CU-COMSEM	0.72	0.88	**0.78**
IRST-BP	0.75	0.80	0.75
ONE-IN-ONE	*1.00*	*0.47*	*0.61*
AUG	0.50	0.88	0.60
PR-cos	0.87	0.47	**0.58**
tf*idf-euc	0.82	0.52	0.57
bin-cos	0.82	0.51	0.56
UBC-AS	0.30	0.91	**0.40**
ALL-IN-ONE	*0.29*	*1.00*	*0.40*

Table 1. Comparative performancegiven among Domain Information and WePS participants

The f-score results show that the difference between our best-performing system PR-cos and the CU-COMSEM and IRST-BP participants is around 20 and 15%. Compared to UBC-AS and ALL-IN-ONE approaches, PR-cos f-score is with 18% better. In comparison to WePS approaches, our domain information approach has the advantage to provide semantic explanation for the groupings following the topic information the web pages carry out.

However, the performance of the two best WePS systems demonstrates that in the future we need to integrate knowledge from the urls, the web page titles and even whole phrases. We have also observed that the systems considering NE information reported better performance. However, in our domain approach we could not map the proper names with the domain information because such evidence is not indicated in *WordNet Domains*.

From the three different domain ranking approaches, the best performing is PR with the cosine similarity. This shows that the reduction of the number of

senses using the domains is useful. In the future, we plan to study the influence of PR with more ample set of categories.

6 Conclusions

In this paper, we have presented a graph-based approach for synset ranking given a domain. The algorithm can be adapted easily to a bigger (smaller) set of categories and is flexible to future change in the number of synsets. PageRank representation shows the communication flow among the synsets and the domains.

We have demonstrated the usage of the synset-domain ranking for the resolution of the WebPS task. We have proposed three different domain ranking methods based on tf*idf, binary vector and PageRank. In order to measure the similarity between the web pages, we have employed the cosine and Euclidean distance measures. Our motivation for the usage of the domain ranking is that two documents are probable to refer to the same individual when the majority of their domains overlaps.

We have evaluated our approach with the WePS data set and we have compared the performance of our systems with WePS participants. The system ranked on 8th and 9th position from sixteen participants.

Finally, the results show that due to the varying web page contents, the domain vector space contains many zero values for shorter documents. Therefore, we want to study the amount of words necessary for the representation of the domain information in a document.

Currently, we did not take advantage of domain proximity such as `MUSIC` and `ENTERTAINMENT`, but we plan to integrate this information as well. Analysing the performances of the WePS systems proposes that named entities, urls, web page titles and noun phrases play important role.

Acknowledgements

This research has been funded by the European Union project QALLME FP6 IST-033860, Spanish Ministry of Science and Technology TEXT-MESS TIN2006-15265-C06-01 and Fundação para a Ciência e a Tecnologia (FCT) scholarship: SFRH / BD / 19909 / 2004.

References

1. Eneko Agirre and Aitor Soroa. Ubc-as: A graph based unsupervised system for induction and classification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 346–349, June 2007.
2. Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 79–85, 1998.

3. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
4. Ying Chen and James H. Martin. Cu-comsem: Exploring rich features for unsupervised web personal name disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 125–128, June 2007.
5. G. Cleuziou, L. Martin, and C. Vrain. Poboc: an overlapping clustering algorithm. application to rule-based classification and textual data. pages 440–444, August 22-27 2004.
6. Andrea Esuli and Fabrizio Sebastiani. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, 2007.
7. Artiles Javier, Julio Gonzalo, and Satoshi Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, 2007.
8. Zornitsa Kozareva, Sonia Vázquez, and Andrés Montoyo. Discovering the underlying meanings and categories of a name through domain and semantic information. In *Proceedings of the Conference on Recent Advances in Natural Language Processing RANLP*, 2007.
9. Zornitsa Kozareva, Sonia Vázquez, and Andrés Montoyo. Multilingual name disambiguation with semantic information. In *TSD*, pages 23–30, 2007.
10. Hugo Liu. Montylingua: An end-to-end natural language processor with common sense. available at: http://web.media.mit.edu/~hugo/montylingua, 2004.
11. Bernardo Magnini and Gabriela Cavaglia. Integrating subject field codes into wordnet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, pages 1413–1418, 2000.
12. Ted Pedersen and Anagha Kulkarni. Unsupervised discrimination of person names in web contexts. In *CICLing*, pages 299–310, 2007.
13. Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. Name discrimination by clustering similar contexts. In *CICLing*, pages 226–237, 2005.
14. Octavian Popescu and Bernardo Magnini. Irst-bp: Web people search using name entities. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 195–198, June 2007.