Researcher affiliation extraction from homepages

István Nagy[1], **Richárd Farkas**[1,2], **Márk Jelasity**[2]
Nagy.Istvan@gmail.com, {rfarkas,jelasity}@inf.u-szeged.hu
[1] University of Szeged, Department of Informatics
Árpad tér 2., H-6720 Szeged, Hungary
[2] Hungarian Academy of Sciences, Research Group on Artificial Intelligence
Aradi vértanuk tere 1., H-6720 Szeged, Hungary

Abstract

Our paper discusses the potential use of Web Content Mining techniques for gathering scientific social information from the homepages of researchers. We will introduce our system which seeks [*affiliation, position, start year, end year*] information tuples on these homepages along with preliminary experimental results. We believe that the lessons learnt from these experiments may be useful for further scientific social web mining.

1 Introduction

Scientific social network analysis (Yang et al., 2009; Said et al., 2008) seeks to discover global patterns in the network of researchers working in a particular field. Common approaches uses bibliographic/scholarly data as the basis for this analysis. In this paper, we will discuss the potential of exploiting other resources as an information source, such as the homepages of researchers. The homepage of a researcher contains several useful pieces of scientific social information like the name of their supervisor, affiliations, academic ranking and so on.

The information on homepages may be present in a structured or natural text form. Here we shall focus on the detection and analysis of full text regions of the homepages as they may contain a huge amount of information while requires more sophisticated analysis than that for structured ones. We will show that this kind of Web-based Relation Extraction requires different techniques than the state-of-the-art seed-based approaches as it has to acquire information from the long-tail of the World Wide Web.

As a case study, we chose one particular scientific social information type and sought to extract information tuples concerning the previous and current *affiliation*s of the researcher in question. We defined 'affiliation' as the current and previous physical workplaces and higher educational institutes of the researcher in question. Our aim is to use this kind of information to discover collegial relationships and workplace-changing behaviour which may be complementary to the items of information originating from bibliographic databases.

Based on a manually annotated corpus we carried out several information extraction experiments. The architecture of the complex system and the recognised problems will be discussed in Section 3, while our empirical results will be presented in Section 4. In the last two sections we will briefly discuss our results and then draw our main conclusions.

2 Related work

The relationship to previous studies will be discussed from a scientific social network analysis as an application point of view and from a Web Content Mining point of view as well.

2.1 Researcher affiliation extraction

Scientific social network analysis has become a growing area in recent years ((Yang et al., 2009; Robardet and Fleury, 2009; Said et al., 2008) just to name a few in recent studies). Its goal is to provide a deeper insight into a research field or into the personal connections among fields by analysing relationships among researchers. The existing studies use the co-authorship (e.g. (Newman, 2001; Barabási et al., 2002)) or/and the citation (Goodrum et al., 2001; Teufel et al., 2006) information – generally by constructing a graph with nodes representing researchers – as the basis for their investigations.

Apart from publication-related relationships – which are presented in structured scholarly datasets –, useful scientific social information can

Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 1–9,
Suntec, Singapore, 7 August 2009.

be gathered from the WWW. Take, for instance the homepage of a researchers where they summarise their *topic of interest*, list *supervisors* and *students*, *nationality*, *age*, *memberships* and so on. Our goal is to develop an automatic Web Content Mining system which crawls the homepages of researchers and extracts useful social information from them.

A case study will be outlined here, where the previous and current affiliations of the researcher in question were gathered automatically. Having a list of normalised *affiliation*s for each researcher of a field (i) we ought to be able to discover collegial relationships (whether they worked with the same group at the same time) which may differ from the co-authorship relation and (ii) we hope to be able to answer questions like *'Do American or European researchers change their workplace more often?'*.

2.2 Information extraction from homepages

From a technology point of view our procedure is a Web Content Mining tool, but it differs from the popular techniques used nowadays. The aim of Web Content Mining (Liu and Chen-Chuan-Chang, 2004) is to extract useful information from the natural language-written parts of websites.

The first attempts on Web Content Mining began with the Internet around 1998-'99 (Adelberg, 1998; Califf and Mooney, 1999; Freitag, 1998; Kosala and Blockeel, 2000). They were expert systems with hand-crafted rules or induced rules used in a supervised manner and based on labeled corpora.

The next generation of approaches on the other hand work in weakly-supervised settings (Etzioni et al., 2005; Sekine, 2006; Bellare et al., 2007). Here, the input is a seed list of target information pairs and the goal is to gather a set of pairs which are related to each other in the same manner as the seed pairs. These pairs may contain related entities (for example, *country - capital city* in (Etzioni et al., 2005) and celebrity partnerships in (Cheng et al., 2009)) or form an entity-attribute pair (like *Nobel Prize recipient - year* in (Feiyu Xu, 2007)) or may be concerned with retrieving all available attributes for entities (Bellare et al., 2007; Paşca, 2009). These systems generally download web pages which contain the seed pairs then learn syntactical/semantical rules from the sentences of the pairs (they generally use the positive instances for one case as negative instances for another case). According to these patterns, they can download a new set of web pages and parse them to acquire new pairs.

These seed-based systems exploit the redundancy of the WWW. They are based on the hypothesis that important information can be found at several places and in several forms on the Web, hence a few accurate rules can be used to collect the required lists. Their goal is to find and recognise (at least) one occurrence of the target information and not to find their every occurrence on the Web. But this is not the case in our scenario. Several pieces of social information for the researchers are available just on their homepages (or nowhere). Thus here we must capture each mention of the information. The weakly-supervised (redundancy-based) systems can build on high-precision and lower recall information extraction, while we have to have target a perfect recall. For the evaluation of such a system we constructed a manually annotated corpus of researchers' homepages. This corpus was also used as a training corpus for the preliminary information extraction experiments described in this paper.

3 The architecture of the system

The general task of our system is to gather scientific social information from the homepages of researchers. In the use case presented in this paper, the input is a set of researchers' names who work in a particular research field (later on, this list can be automatically gathered, for example, from a call for papers) and the output is a list of affiliations for each researcher. Here the affiliation is a tuple of *affiliation*, *position type* and *start/end dates*. We think that the lessons learnt from affiliation extraction will be useful for the development of a general social information extraction system.

The system has to solve several subproblems which will be described in the following subsections.

3.1 Locating the homepage of the researcher

Homepage candidates can be efficiently found by using *web search engine* queries for the given name. In our case study the homepage of the researcher (when it existed) were among the top 10 responses of the Google API[1] in each case. However, selecting the correct homepage from the top 10 responses is a harder task. Among

[1] http://code.google.com/apis/

these sites there are (i) publication-related ones (books/articles written by the researchers, call for papers), sites of the institute/group associated with the researcher and (ii) homepages of people sharing the same name.

In our preliminary experiments, we ignored these two basic problems and automatically parsed each website. However in the future we plan to develop a two-stage approach to solve them. In the first stage a general homepage detection model – a binary classification problem with classes `homepage`/`non-homepage` – will be applied. In the second stage we will attempt to automatically extract textual clues for the relations among the researchers (e.g. the particular field they work in) from the homepage candidates and utilise these cues for name disambiguation along with other biographical cues. For a survey of state-of-the-art name disambiguation, see (Artiles et al., 2009).

3.2 Locating the relevant parts of the site

The URL got from the search engine usually points to the main page of the homepage site. An ideal system should automatically find every page which might contain scientific social information like *Curriculum Vitae*, *Research interests*, *Projects* etc. This can be done by analysing the text of the links or even the linked page. In our case study we simply parsed the pages to a depth of 1 (i.e. the main page and each page which was linked from it).

The located web pages usually have their content arranged in sections. The first step of information extraction may be a relevant section selection module. For example, in the affiliation extraction task the *Positions Held* and *Education* type sections are relevant while *Selected Papers* is not. Having several relevant sections with their textual positions, an automatic classification system can filter out a huge number of probably irrelevant sections. In our experiments, we statistically collected a few "relevant keywords" and filtered out sections and paragraphs which did not contain any of these keywords.

3.3 Extracting information tuples

Pieces of scientific social information are usually present on the homepages and in the CVs even in an itemised (structured) form or in a natural language full text form. Information extraction is performed from the structured parts of the documents by automatically constructed rules based on the HTML tags and keywords. This field is called Wrapper Induction (Kushmerick, 2000).

We shall focus on the information extraction from raw texts here because we found that more pages express content in textual form than in a structured one in the researchers' homepages of our case study and this task still has several unsolved problems. We mentioned above that scientific social information extraction has to capture each occurrence of the target information. We manually labeled homepages for the evaluation of these systems. We think that the DOM structure of the homepages (e.g. formatting tags, section headers) could provide useful information, hence the labeling was carried out in their original HTML form (Farkas et al., 2008). In our preliminary experiments we also used this corpus to train classification models (they were evaluated in a one-researcher-leave-out scheme). The purpose of these supervised experiments was to gain an insight into the nature of the problem, but we suggest that a real-world system for this task should work in a weakly-supervised setting.

3.4 Normalisation

The output of the extraction phase outlined above is a list of *affiliation*s for each researcher in the form that occurred in the documents. However, for scientific social network analysis, several normalisation steps should be performed. For example, for collegial relationship extraction, along with the matching of various transliteration of research groups (like *Massachusetts Institute of Technology* and *MIT AI Lab*), we have to identify the appropriate institutional level where two researchers probably still have a personal contact as well.

4 Experiments

Now we will present the affiliation corpus which was constructed manually for evaluation purposes along with several preliminary experiments on affiliation extraction.

4.1 The affiliation corpus

We manually constructed a web page corpus containing HTML documents annotated for publicly available information about researchers. We downloaded 455 sites, 5282 pages for 89 researchers (who form the Programme Committee of the SASO07 conference[2]), and two indepen-

[2] http://projects.csail.mit.edu/saso2007/tmc.html

dent annotators carried out their manual labeling in the original (HTML) format of the web pages, following an annotation guideline (Farkas et al., 2008). All the labels that were judged inconsistent were collected together from the corpus for a review by the two annotators and the chief annotator. We defined a three-level deep annotation hierarchy with 44 classes (labels). The wide range of the labels and the inter-annotator agreement both suggest that the automatic reproduction of this full labelling is a hard task.

We selected one particular information class, namely *affiliation* from our class hierarchy for our case study. We defined 'affiliation' as the current and previous physical workplaces and higher educational institutes of the researcher in question as we would like to use this kind of information to discover collegial relationships and workplace-changing behaviour. Here institutes related to review activities, awards, or memberships are not regarded as affiliations. We call `position` the tuple of <`affiliation`, `position_types`, `years`>, as for example in <*National Department of Computer Science and Operational Research at the University of Montreal*, *adjunct Professor*, {*1995*, *2002*}>[3]. Among the four slots just the `affiliation` slot is mandatory (it is the head) as the others are usually missing in real homepages.

The problem of finding the relevant pages of a homepage site originating from a seed URL was not addressed in this study. We found that pages holding affiliation information was the one retrieved by Google in 135 cases and directly linked to the main page in 50 cases. We found affiliation information for all of the 89 researchers of our case study in the depth of 1, but we did not check whether deeper crawling could have yielded new information.

The affiliation information (like every piece of scientific social information) can be present on web pages in an itemised or natural text format. We manually investigated our corpus and found that the 47% of the pages contained affiliation information exclusively in a textual form, 24% exclusively in an itemised form and 29% were hybrid. Information extraction from these two formats requires different methods. We decided to address the problem of affiliation extraction just by using the raw text parts of the homepages.

We partitioned each downloaded page at HTML breaking tags and kept the parts (paragraphs) which were regarded as "raw text". Here we used the following rule: *a textual paragraph has to be longer than 40 characters and contain at least one verb*. Certainly this rule is far from perfect (paragraphs describing publication and longer items of lists are still present), but it seems to be a reasonable one as it extracts paragraphs even from 'hybrid' pages. We found 86,735 paragraphs in the 5282 downloaded pages and used them in experiments in a raw `txt` format (HTML tags were removed).

Table 4.1 summarises the size-related figures for the part of this textual corpus which contains affiliation information (these paragraphs contain manually labeled information). The corpus is freely available for non-commercial use[4].

# researchers	59
# pages	103
# paragraph	151
# sentences	181
# `affiliation`	374
# `position_type`	326
# `year`	212

Table 1: The size of the textual corpus which contains affiliation information.

4.2 The multi-stage model of relation extraction

Our relation extraction system follows the architecture described in the previous section. We focus on the *relevant part location* and *information extraction* steps in this study. We applied simple rules to recognise the relevant parts of the homepages. We extract textual paragraphs as described above and then filter out probably irrelevant ones (Section 4.3).

Preliminary supervised information extraction experiments were carried out in our case study in order to get an insight into the special nature of the problem. We used a one-researcher-leave-out evaluation setting (i.e. the train sets consisted of the paragraphs of 88 researchers and the test sets concerned 1 researcher), thus we avoided the situations where a training set contained possibly re-

[3] the example is extracted from http://bcr2.uwaterloo.ca/∼rboutaba/biography2.htm

[4] www.inf.u-szeged.hu/rgai/homepagecorpus

dundant information about the subject of the test texts.

A two-stage information extraction system was applied here. In the first phase, a model should recognise each possible slot/entities of the target information tuples (Section 4.4). Then the tuples have to be filled, i.e. the roles have to be assigned and irrelevant entities should be ignored (Section 4.5).

4.3 Paragraph filtering

Because just a small portion of extracted textual paragraphs contained affiliation information, we carried out experiments on filtering out probably irrelevant paragraphs.

Our filtering method exploited the paragraphs containing `position` (positive paragraphs). We calculated the $P(word|positive)$ conditional probabilities and the best words based on this measure (e.g. *university*, *institute* and *professor*) then formed the so-called positive wordset. The paragraphs which did not contain any word from the positive wordset were removed. Note that standard positive and negative sample-based classification is not applicable here as the non-positive paragraphs may contain these indicative words, but in an irrelevant context or with a connection to people outside of our scope of interest. Our 1-DNF hypothesis described above uses just positive examples and it was inspired by (Yu et al., 2002).

After performing this procedure we kept 14,686 paragraphs (from the full set of 86,735), but we did not leave out any annotated text. Hence the information extraction module could then work with a smaller and less noisy dataset.

4.4 Detecting possible slots

We investigated a Named Entity Recognition (NER) tool for detecting possible actors of a `position` tuple. But note that this task is not a classical NER problem because our goal here is to recognise just those entities which may play a role in a `position` event. For example there were many `year` tokens in the text – having the same orthographic properties – but only a few were related to affiliation information. The contexts of the tokens should play an important role in this kind of an NER targeting of very narrow semantic NE classes.

For training and evaluating the NER systems, we used each 151 paragraphs containing at least one manually labeled `position` along with 200 other manually selected paragraphs which do not contain any labeled `position`. We decided to use just this 151+200 paragraphs instead of the full set of 86,735 paragraphs for CPU time reasons. Manual selection – instead of random sampling – was required as there were several paragraphs which contained affiliation information unrelated to the researcher in question, thus introducing noise. In our multi-stage architecture, the NER model trained on this reduced document set was than predicated for the full set of paragraphs and false positives (note that the paragraphs outside the NER-train do not contain any gold-standard annotation) has to be eliminated.

We employed the Condition Random Fields (Lafferty et al., 2001) (implementation MALLET (McCallum, 2002)) for our NER experiments. The feature set employed was developed for general NER and includes the following categories (Szarvas et al., 2006):

orthographical features: capitalisation, word length, bit information about the word form (contains a digit or not, has uppercase character inside the word, and so on), character level bi/trigrams,

dictionaries of first names, company types, denominators of locations,

frequency information: frequency of the token, the ratio of the token's capitalised and lowercase occurrences, the ratio of capitalised and sentence beginning frequencies of the token which was derived from the Gigaword dataset[5],

contextual information: sentence position, trigger words (the most frequent and unambiguous tokens in a window around the NEs) from the train text, the word between quotes, and so on.

This basic set was extended by two domain-specific gazetteers, namely a list of *university names* and *position types*. We should add that a domain-specific exception list (containing e.g. *Dr.*, *Ph.D.*) for augmenting a general sentence splitter was employed here.

Table 2 lists the phrase-level $F_{\beta=1}$ results obtained by CRF in the one-researcher-leave-out

[5]Linguistic Data Consortium (LDC), catalogId: LDC2003T05

evaluation scheme, while Table 3 lists the results of a baseline method which labels each member of the university and position type gazetteers and identifies years using regular expressions. This comparison highlights the fact that labeling each occurrences of this easily recognisable classes cannot be applied. It gives an extremely low precision thus contextual information has to be leveraged.

	Precision	Recall	$F_{\beta=1}$
affiliation	66.78	53.28	59.27
position type	87.50	70.22	77.91
year	86.42	69.31	76.92
TOTAL	78.73	62.88	69.92

Table 2: The results achieved by CRF.

	Precision	Recall	$F_{\beta=1}$
affiliation	21.43	9.68	13.33
position type	23.27	66.77	34.51
year	65.77	98.99	79.03
TOTAL	32.16	44.08	37.19

Table 3: NER baseline results.

4.5 The assignment of roles

When we apply the NER module to unknown documents we have to decide whether the recognised entities have any connection with the particular person as downloaded pages often contain information about other researchers (supervisors, students, etc.) as well. The subject of the information is generally expressed by a proper noun at the beginning of the page or paragraph and then anaphoric references are used. We assumed here that each `position` tuple in a paragraph was related to exactly one person and when the subject of the first sentence of the paragraph was a personal pronoun *I*, *she*, *he* then the paragraph belonged to the author of the page.

To automatically find the subject of the paragraphs we tried out two procedures and evaluated them on the predictions of the NER model introduced in the previous subsection. First, we applied a NER trained on the person names of the CoNLL-2003 corpus (Tjong Kim Sang and De Meulder, 2003). The names predicted by this method were then compared to the owner of the homepage using name normalisation techniques. If no name was found by the tagger we regarded the paragraph as belonging to the author. Its errors had two sources; the NER trained on an out-domain corpus made a lot of false negatives and the normalisation method had to deal with incorrect "names" (like *Paul Hunter Curator* as a name phrase) as well.

The second method was simpler. We kept the `position` tuples whose paragraph contained any part of the researcher name or any of the "*I*", "*she*", "*he*" personal pronouns. Its errors came, for instance, from finding the "*Paul*" string for "*Paul Robertson*" in the text snippet "*Paul Berger*".

We applied these two subject detection methods to the predictions of our slot detection NER modul. Table 4 summarises the accuracies of the systems, i.e. whether they made the correct decision on "is this forecasted affiliation corresponds to the researcher in question". The columns of this table shows how many `affiliation` prediction was carried out by the slot detection system, i.e. how many times has to made a decision. "name. det" and "p. pronouns" refer to the two methods, to the name detection-based and to the personal pronoun-matcher ones. We investigated their performance on the paragraphs which contained manually labeled information, on the paragraphs which did contained any but the slot detection module forecasted at least one `affiliation` here and on the union of these sets of paragraphs. The figures of the table shows that the personal pronoun detection approach performs significantly better on the paragraphs which really contains affiliation information. This is due to the fact that this method removes less prediction compared to the name based one and there are just a few forecast which has to be removed on the paragraphs which contain information.

	#pred	name det.	p. pronouns
annotated	165	66.9	87.8
non-ann.	214	71.5	61.2
full set	379	69.4	73.4

Table 4: Accuracies of subject detection methods.

To find relationships among the other types of predicated entities (*affiliation, position type, start year, end year*) we used a very simple heuristic. As the `affiliation` slot is the head of the tuple we simply assigned every other detected entity to the nearest `affiliation` and regarded the earlier preidcated `year` token as the `start year`.

This method made the correct decision in the 91.3% and 71.8% of the cases applied on the gold-standard annotation and the predicated entities, respectively. We should add that using the predicted labels during the evaluation, the false positives of the NER counts automatically an error in relation detection as well.

5 Discussion

The first step of the information extraction system of this case study was the localisation of relevant information. We found that Web search engines are efficient tools for finding homepages. We empirically showed that a very simple crawling (downloading everything to a depth of 1) can be applied, because the irrelevant contents can be removed later. The advantage of focused crawling (i.e. making a decision before downloading a linked page) is that it can avoid the time-consuming analysis of pages. However making the decision of whether the linked document might contain relevant information is a hard task. On the other hand we showed that the requested information is reachable in depth 1 and that a fast string-matching based filtering method can significantly reduce the amount of texts which have to be analysed without losing any information. Moreover, the positive example-based filtering approach can be employed in a seed-driven setting as well.

For the information extraction phase we think that a high-recall system has to be developed. We constructed a corpus with contextual occurrences for evaluation issues. The extraction can be relationship detection-based (e.g. the state-of-the-art seed-driven approaches seek to acquire syntactic/semantic patterns which are typical of the relationship itself) or entity-based (like our method, these approaches first identify possible actors then look for relationships among them). We expect that the latter one is more suitable for high-recall tasks.

The NER system of this case study achieved significantly better results than those for the baseline method. We experimentally showed that it could exploit the contextual information and that the labeled entities were those which were affiliation-related. However, the overall system has to be improved in the future. We manually analysed the errors on a part of the corpus and found a few typical errors were present. Our annotation guide said that the geographical location of the affiliation was a part of the affiliation as it sometimes identifies the department (e.g. *"Hewlett-Packard Labs in Palo Alto"*). This extension of the phrase proved to be difficult because there were several cases with the same orthographic features (e.g. *Ph.D. from MIT in Physics*). The acronyms immediately after the affiliation are a similar case, which we regard as part of the name and it is difficult for the NER to handle (e.g. *Centre for Policy Modelling (CPM)*). As there is no partial credit; an incorrect entity boundary is penalised both as a false positive and as a false negative.

These points also explain the surprisingly low precision of the baseline system as it labeled university names without more detailed identification of the unit (e.g. *Department of Computer Science, [Waterloo University]*$_{BASELINE}$). We should add that these two annotation guidelines are questionable, but we expect that information might get lost without them. Moreover, there is an another reason for the low recall, it is that our human annotators found textual clues for *position type*s on verbs as well (e.g. *I lead*$_{TYPE}$ *the Distributed Systems Group*). The context of these labeled examples are clearly different from that of the usual *position type*.

Comparing the two subject detection methods, we see that the name detection model which learnt on an out-domain corpus made a lot of mistakes, thus the method based on it judged more paragraphs as irrelevant ones. The name detection could be improved by a domain corpus (for example the training corpus did not contain any *Prof. NAME* example) and by applying more sophisticated name normalisation techniques. When we manually analysed the errors of these procedures we found that each false negative of the simpler subject detection method was due to the errors of the textual paragraph identification definition used. There were several itemisations whose header was type of *"Previously I worked for:"* and the textual items themselves did not contain the subject of the affiliation information. The false positives often originated from pages which did not belong to the researcher in question but contained him name (e.g. *I am a Ph.D. Student working under the supervision of Prof. NAME*).

Lastly, an error analysis of the affiliation head seeking heuristic revealed that the 44% of the predicted `position type` and `year` entities's

sentences did not contain any `affiliation` prediction. With the gold-standard labeling, there were 6 sentences without `affiliation` labels and only one of them used an anaphoric reference, the others were a consequence of the erroneous automatic sentence splitting of the HTML documents. The prediction of the NER system contained many more sentences without any `affiliation` label. These could be fixed by forcing a second forecast phase to predict `affiliation` in these sentences or by removing these labels in a post-processing step.

The remaining errors of the affiliation head assignment could be avoided just by employing a proper syntactic analyser. The most important linguistic phenomena which should be automatically identify for this problem is enumeration. For instance, we should distinguish between the enumeration and clause splitting roles of 'and' (e.g. *"I'm a senior researcher and leader of the GROUP"* and *"He got his PhD from UNIVERSITY1 in YEAR and has a Masters from UNIVERSITY2"*). This requires a deep syntactic analysis, i.e. the use of a dependency parser which has to make accurate predictions on several certain types of dependencies is probably needed.

6 Conclusions

In this paper we introduced a Web Content Mining system for gathering affiliation information from the homepages of researchers. The affiliation information collected from this source might be of great value for scientific social network analysis.

We discussed the special nature of this task compared to common Web-based relation extraction approaches and identified several subtasks of the system during our preliminary experiments. We argued that the evaluation of this kind of system should be carried out on a manually labeled reference corpus. We introduced simple but effective solutions for the subproblems along with empirical results on a corpus. We achieved reasonable results with an overall phrase-level $F_{\beta=1}$ score of 70% on the possible slot detection and an accuracy of 61% on relation extraction (as an aggregation of the subject detection and the affiliation head selection procedures). However each subproblem requires more sophisticated solutions, which we plan to address in the near future.

Acknowledgments

This work was supported in part by the NKTH grant of the Jedlik Ányos R&D Programme (project codename TEXTREND) of the Hungarian government. The authors would like to thank the annotators of the corpus for their devoted efforts.

References

Brad Adelberg. 1998. Nodose - a tool for semi-automatically extracting structured and semistructured data from text documents. *ACM SIGMOD*, 27(2):283–294.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009. Weps 2 evaluation campaign: overview of the web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614.

Kedar Bellare, Partha Talukdar, Giridhar Kumaran, Fernando Pereira, Mark Liberman, Andrew McCallum, and Mark Dredze. 2007. Lightly-supervised attribute extraction for web search. In *Proceedings of NIPS 2007 Workshop on Machine Learning for Web Search*.

Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 328–334.

Xiwen Cheng, Peter Adolphs, Feiyu Xu, Hans Uszkoreit, and Hong Li. 2009. Gossip galore – a self-learning agent for exchanging pop trivia. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 13–16, Athens, Greece, April. Association for Computational Linguistics.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana maria Popescu, Tal Shaked, Stephen Soderl, Daniel S. Weld, and Er Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165:91–134.

Richárd Farkas, Róbert Ormándi, Márk Jelasity, and János Csirik. 2008. A manually annotated html corpus for a novel scientific trend analysis. In *Proc. of The Eighth IAPR Workshop on Document Analysis Systems*.

Hong Li Feiyu Xu, Hans Uszkoreit. 2007. A seed-driven bottom-up machine learning framework for

extracting relations of various complexity. In *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 6.

Dayne Freitag. 1998. Information extraction from html: Application of a general machine learning approach. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 517–523.

A. A Goodrum, K. W McCain, S. Lawrence, and C. L Giles. 2001. Scholarly publishing in the internet age: a citation analysis of computer science literature. *Information Processing and Management*, 37:661–675, September.

Raymond Kosala and Hendrik Blockeel. 2000. Web mining research: A survey. *SIGKDD Explorations*, 2:1–15.

Nicholas Kushmerick. 2000. Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118:2000.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Bing Liu and Kevin Chen-Chuan-Chang. 2004. Editorial: special issue on web content mining. *SIGKDD Explor. Newsl.*, 6(2):1–4.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

M. E. J. Newman. 2001. The structure of scientific collaboration networks. In *Proceedings National Academy of Sciences USA*, pages 404–418.

Marius Paşca. 2009. Outclassing Wikipedia in open-domain information extraction: Weakly-supervised acquisition of attributes over conceptual hierarchies. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, March.

Celine Robardet and Eric Fleury. 2009. Communities detection and the analysis of their dynamics in collaborative networks. *Int. J. Web Based Communities*, 5(2):195–211.

Yasmin H. Said, Edward J. Wegman, Walid K. Sharabati, and John T. Rigsby. 2008. Social networks of author-coauthor relationships. *Computational Statistics & Data Analysis*, 52(4):2177–2184.

Satoshi Sekine. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 731–738, Sydney, Australia, July. Association for Computational Linguistics.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. *DS2006, LNAI*, 4265:267–278.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87, Sydney, Australia, July. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Y. Yang, C. M. Au Yeung, M. J. Weal, and H. Davis. 2009. The researcher social network: A social network based on metadata of scientific publications.

Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. 2002. Pebl: positive example based learning for web page classification using svm. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248, New York, NY, USA. ACM.