A manually annotated HTML corpus for a novel scientific trend analysis

Richárd Farkas[1], Róbert Ormándi[1], Márk Jelasity[1] and János Csirik[2]
[1] *MTA-SZTE, Research Group on Artificial Intelligence,*
6720 Szeged, Aradi Vértanúk tere 1., Hungary,
[2] *University of Szeged, Department of Informatics,*
6720 Szeged, Árpád tér 2., Hungary,
{rfarkas,ormandi,jelasity,csirik}@inf.u-szeged.hu

Abstract

Here we present a manually annotated corpus of web pages and annotation tool for Web Content Mining. The corpus is extensively annotated, has a hierarchical label structure and is freely available for research purposes. The annotation tool is a Firefox extension which allows the annotator to work with the pages in their original appearance. This tool handles the annotation hierarchy independently of the DOM tree of the web pages, and it allows overlapped annotation between the HTML tags.

1. Introduction

Because of the Internet and the globalisation process the amount of available information is growing at an incredible rate. The greatest part of this information is in textual form, usually in web pages that are designed for humans to read. This amount of information requires the use of a computer in processing tasks, like IE and document classification tasks.

The aim of Web Content Mining [11] is to extract useful information from the natural language-written parts of websites. After several early attempts on Web Content Mining in the late nineties, the researchers of the Web Mining community now focus on Web Usage Mining [4][13][18] and Wrapper Induction [10]. The Web Usage Mining refers to the automatic discovery and analysis of patterns in data collected or generated via result of user interactions with Web resources on one or more websites. One goal is to capture and analyze the behavioural patterns and profiles of users interacting with a website. Another goal of Wrapper Induction is the automatic extraction of information from structured documents like product information from webshops. On the other hand the Computation Linguistic community focuses on the raw text, i.e. it rarely deals with the structural information of the documents. We think that, based on the recent improvements in statistical Natural Language Processing, Web Content Mining will become a rapidly emerging area over the coming years. The demand for it is growing because there are views that the real-world Semantic Web could not come into existence without the automatic content handling of web pages.

We introduce here a manually and extensively annotated corpus for Web Content Mining. It is freely available for research purposes and we hope it will become a basis for the validation and comparison of automatic Web Content Mining systems.

The domain that the corpus covers can be grouped into social networks analysis [3][8][12] and scientific trend monitoring [14], which have become growing areas in recent years. Our main interest here is to investigate the utility of information besides publication and citation data – such as the *student-supervisor* relationship, self-written *research interest* or *programme committee memberships* – for assessing scientific communities and trends. The corpus described in this article contains the detailed manual annotation of homepages of researchers.

Our annotation method has several levels and aims. It includes document level tags (like *this HTML is not relevant* or *the picture is the portrait of the researcher*), layout information (such as which part of the HTML site deals with the previous workplaces of the researcher) and fine-grained information slots (like the *year* and the *name of courses taught*). We believe that the layout and the structure of these documents – which also applies to medical records and research articles as well – contain a lot of useful information.

Thus our intention was to keep the documents in their original form in the corpus design.

2. Related work

The first attempts on Web Content Mining began with the Internet around '98-'99 [1-2][5][9]. They were expert systems with hand-crafted rules or induced rules used in a supervised manner and based on labelled corpora [1].

These systems (and the corpora) focused on one or two specialized information types. Here we tried to cover the whole range of informative types and we have 44 labels. These early works reported on the utilization of the HTML structure, but they did not usually handle real structures. Instead they considered each tag as one special token and then applied standard Natural Language Processing techniques.

In the past decade, the number of papers on the analysis of semi-structured documents has been quite low (e.g. [7]). There are many articles on Information Extraction [16] which focus on the natural language parts and if the texts come from web pages the tags are removed and there also many articles on Wrapper Induction [10] which focus just on the structure of the documents and do not use, or employ very basic, Natural Language Processing techniques. The automatic labelling of websites with the information hierarchy of our corpus requires the joint application of these two approaches. Think, for example, of the page-long research interests of a researcher's homepage and the enumeration of positions held. To the best of our knowledge there is no other freely available (for non-commercial usage) corpora which contains extensively and manually annotated web pages.

3. The HTML annotation tool

For the efficient manual annotation of the corpus, a user-friendly software tool was needed. We compiled a list of requirements for this annotation tool. They are the following:

- The annotators should work on the pages in their original appearance, hence they should not work on source HTMLs and we should not use labelling which would modify the appearance of a page. Moreover, as the corpus contains downloadable subsites, the tool has to be compatible with the hyperlinks.
- The labelled parts of the document should not match the DOM tree of the page. The libelling (and the tool) has to support the cases which can be seen in Figure 1. If we would like to annotate the teaching activities in the original page (see below), we have to have an overlapped annotation, because the original page is badly structured. So the beginning of the teaching annotation has to start in the middle of the first paragraph tag and has to end in the end of the second paragraph tag. This overlapped annotation can be seen on the RHS of Figure 1.
- Our type-family is hierarchical, hence the tool has to automatically verify the consistency of the annotation hierarchy.

We could not find any off-the-self solution which fulfils all of the above criteria, as WYSIWYG HTML editors and "semantic web annotators" do not provide hierarchical and out-of DOM tagging, and raw text annotators do not handle HTML layout and browsing aspects.

We decided to develop an annotation tool which can be freely downloaded from the following corpus website: http://www.inf.u-szeged.hu/rgai/homepagecorpus. The user can browse the downloaded web pages, select an arbitrary part of the page and get the allowed labels (hierarchically consistent with the already tagged regions) for his selection.

Original page:

```
<P>
CV:<BR/>
Free text CV...<BR/>
Teching: <BR/>
</P>
<UL>
  <LI>Teaching activity #1</LI>
  <LI>Teaching activity #2</LI>
</UL>
<P>
Free text about teaching experiencies...
</P>
```

Annotated page:

```
<P>
CV:<BR/>
Free text CV...<BR/><!--TEACHING-begin-->
Teching: <BR/>
</P>
<UL>
  <LI><!--TEACHING_COURSE-begin-->Teaching activity #1<!--TEACHING_COURSE-end--></LI>
  <LI><!--TEACHING_COURSE-begin-->Teaching activity #2<!--TEACHING_COURSE-end--></LI>
</UL>
<P>
Free text about teaching experiencies...<!--TEACHING-end-->
</P>
```

Figure 1. Overlapped annotation

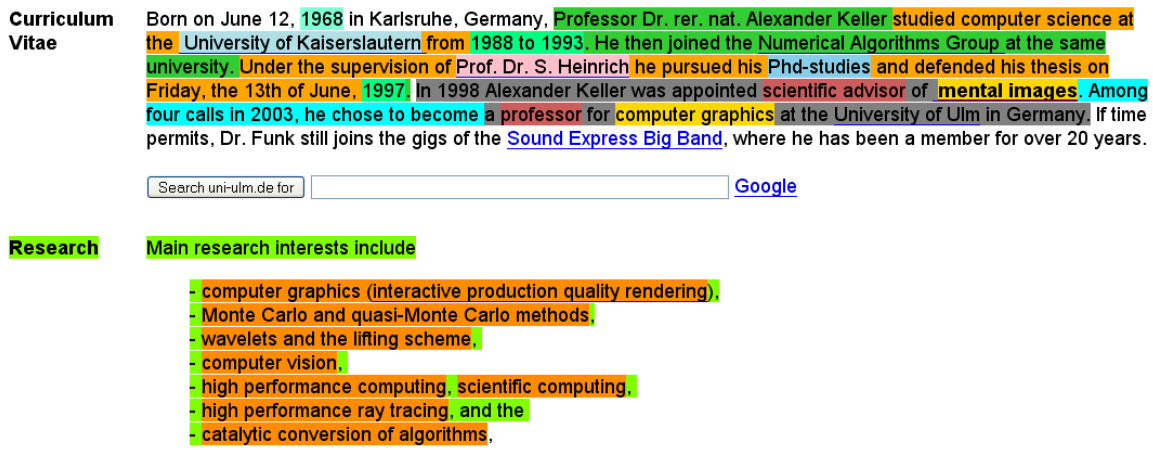

Figure 2. A screenshot of the annotation

The tool places a special HTML comment tag at the beginning and end of the selection. The use of comment tags rather than some other kind of HTML tag helps preserve the page's appearance and provides out-of DOM labelling. The tool is a Firefox extension; hence its installation and usage are both very simple.

4. The corpus

In this section we will elaborate on the corpus we developed. It is freely accessible for academic purposes at: http://inf.u-szeged.hu/rgai/homepagecorpus.

4.1. Obtaining the documents

We decided to use the programme committee of the First IEEE International Conference on Self-Adaptive and Self-Organizing Systems 2007[1] conference as the raw material for our corpus. We obtained the names of committee members from the conference web pages and performed Google queries on them using the Google Search API. Then the union of the top 10 Google result URLs formed the seed set of crawling. To avoid the overloading of the servers we just downloaded neighbouring seed pages: we downloaded only HTMLs and image files which were linked to the seed page.

First we evaluated this very simple "focused crawling" approach and asked our annotators to classify each seed as *homepage* (the professional homepage of the researcher), *relevant* or *non-relevant.* We found that this simple method located the homepage of a given researcher in roughly 12.77%[2] of the cases.

4.2. The label hierarchy

We defined a three-level deep annotation hierarchy with 44 labels. Here the root level annotations represent layout-type labelling. For example, in Figure 2 the enumeration of '*research interest*'-s and the corresponding header belong to one root level annotation and the first marked in green and the HTML comment-pair:

```
<!--RESEARCH_INTERESTS-begin-->
<!--RESEARCH_INTERESTS-end-->
```

The second level of the annotation hierarchy represents one particular event or piece of information. It contains the deepest level labels, which are well-defined slots of information. For example the upper part of the web page shown in Figure 2 – which is actually a natural language paragraph – has a layout level label for '*educations*' (dark-green), which contains two particular abstract entities of the '*education*' label (orange). The second one has three slots (deepest level in the hierarchy): *supervisor* (pink), *degree* (blue) and *year* (green)[3].

4.3. The annotation process

The corpus annotation work is still in progress. Two independent annotators do the manual labelling, following the guidelines written by a senior researcher

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

[2] This is the average recognition score of the two annotators.

[3] In the grey-scale figure, you cannot follow this example.

before the annotation of the corpus was initiated. These guidelines had to be amended several times in the annotation stage as the annotators were often confronted with problematic issues. The annotators are not allowed to communicate with each other as far as the annotation process is concerned. When two given annotations have to be discussed and settled, any differences between the two will be resolved by the senior researcher, yielding the gold standard labelling of the corpus.

4.4. Corpus statistics

The Committee of the First IEEE International Conference on Self-Adaptive and Self-Organizing Systems 2007 conference is made up 89 researchers, which provided the basis for Google queries. Using the Google Search API, we found 455 downloadable URLs. These URLs comprised the seed set of crawling. This means that the average downloadable hits per person was only 5.11, which is quite low. This may be attributed to unwanted query results, like the results of DBLP [4] and other computer science bibliography portals. In the end, we downloaded 5282 HTML files whose average file size was 5.98kB.

Currently the two annotators are working independently on the corpus. Table 1 shows the status of their annotation work and the results on the seed set classification task.

Table 1. A screenshot of the annotation

	Annotator 1	**Annotator 2**
Number of annotated researchers	48	24
Number of annotated files	2813	1242
homepage **annotated rate**	2.85%	22.69%
relevant **annotated rate**	15.45%	31.93%
non-relevant **annotated rate**	31.30%	21.85%

As the reader can see, Annotator 1 has processed much more data than Annotator 2 so far, but his *homepage* and *relevant* rates are much lower, which could mean that his annotation is not so thorough. One can see very big differences between the *homepage* annotated rates of the two annotators, which could cause low inter-annotation agreement in this task.

[4] http://www.informatik.uni-trier.de/~ley/db/

Below tables 2, 3 and 4 show the frequencies of the labels of different annotators and different hierarchical levels.

Table 1 contains the label distribution derived from the first level of the hierarchy. As can be seen, the well-defined labels like CURRENT_POSITION, EDUCATIONS, STUDENTS and TEACHING show a relatively high correlation between the two annotators, but the weakly-defined ones like OTHER_SOCIAL_INFORMATION do not. There is also a big difference in the PROJECTS label, which is a well-defined label as well. This difference and the fact that the Annotator 2 has so far annotated many more labels are surely attributable to Annotator 2's thoroughness and devotion.

Table 2. Label frequency distribution for Level 1

Level 1	**Annotator 1**	**Annotator 2**
BIRTH_YEAR	0	1
CURRENT_POSITION	31	30
CURRENT_POSITION_AFFILIATION	25	17
CURRENT_POSITIONS	0	16
EDUCATIONS	15	18
INVITED_TALKS	3	5
OTHER_SOCIAL_INFORMATION	16	42
PREVIOUS_POSITION_AFFILIATION	0	3
PREVIOUS_POSITIONS	9	19
PROFESSIONAL_MEMBERSHIPS	5	6
PROGRAM_COMITEES	9	14
PROJECTS	17	39
RESEARCH_INTEREST	34	38
REVIEWS	3	0
STUDENTS	6	9
TEACHING	12	12

Table 3 presents the frequencies of labels of the second level. In the second level labelling, we observe a big difference between the two annotators. One part of this difference might be caused by the inheritance of the parent label in the hierarchy; for example, the PROJECT label. The parent of the PROJECT label is PROJECTS. In the previous level, the PROJECTS label had quite different frequencies, which implies that the inherited PROJECT label should be different as well (the PROJECT label cannot be annotated outside the scope of a PROJECTS label).

Table 3. Label frequency distribution for Level 2

Level 2	Annotator 1	Annotator 2
CURRENT_POSITION	0	22
CURRENT_POSITION_YEAR	6	8
EDUCATION	42	34
INVITED_TALK	6	19
MEMBERSHIP_ORGANISATION	12	22
PREVIOUS_POSITION	14	33
PROGRAM_COMITEE	31	55
PROJECT	40	63
REVIEW	7	0
STUDENT_NAME	123	51
TEACHING_COURSE	19	26

Table 4 below gives the distribution scores of the label frequencies of the two annotations.

Table 4. Label frequency distribution for Level 3

Level 3	Annotator 1	Annotator 2
CURRENT_POSITION_AFFILIATION	0	9
CURRENT_POSITION_COLLAUGES	0	1
CURRENT_POSITION_YEAR	0	2
EDUCATION_AFFILIATION	0	15
EDUCATION_DEGREE	21	26
EDUCATION_SUPERVISOR	5	9
EDUCATION_YEAR	19	21
INVITED_TALK_CONFERENCE	4	13
INVITED_TALK_YEAR	6	13
PREVIOUS_POSITION_AFFILIATION	0	8
PREVIOUS_POSITION_YEAR	5	12
PROGRAM_COMITEE_CONFERENCE	27	62
PROGRAM_COMITEE_TYPE	0	16
PROGRAM_COMITEE_YEAR	28	39
PROJECT_COLLAUGES	17	60
PROJECT_NAME	32	96
PROJECT_TOPIC	8	32
REVIEW_JOURNAL	7	0
REVIEW_YEAR	7	0
TEACHING_AFFILIATION	2	0
TEACHING_SUBJECT	19	25
TEACHING_YEAR	13	40

The results of this table are quite similar to those of Table 3. Here we see that some series of frequencies as paths in the labelling tree (e.g. 'EDUCATIONS' – 'EDUCATION' – 'EDUCATION_DEGREE') go together in the case of different annotators.

In summary, we can say that the label frequency distributions of the different levels have big differences, which suggests that the inter-annotation agreement might be poor and may mean that this annotation task is quite hard for human annotators.

In the next section we will put forward some evaluation metrics for our corpus.

5. Evaluation issues

The evaluation of a hierarchical annotation is not trivial, so here we would like to propose some metrics.

Two different annotations for a web page can be viewed as two forests. These forests may be made up of a different number of trees - which makes the evaluation more difficult. To overcome this, we suggest assigning a fictive root node called *#ROOT#* to each forest and adding each tree of forests to the corresponding fictive root node. After doing this, we have to evaluate the two trees in order to evaluate the two different annotations.

At this point, any standard tree similarity measures [15] could be used to evaluate the annotations, but here we recommend two:

- $F_{\beta=1}$-measure [6][17],
- Symmetric tree difference [15].

6. Conclusions

In this paper we reported on the construction of a corpus containing HTML documents annotated for publicly available information about researchers. The corpus is accessible for academic purposes and is free of charge. Apart from the intended goal of serving as a common resource for the testing and comparison of automatic Web Content Mining systems, we think that interesting and novel scientific trend analyses can be carried out based on the corpus.

The wide range of the labels and the inter-annotator agreement suggest that the automatic reproduction of the labelling is a hard task. Furthermore, the agreement rate among humans can be regarded as the theoretical upper limit for an automatic system.

The moderate size of the corpus and the large number of labels do not permit the training of supervised Web Information Extraction systems. However it is a good candidate for the construction and evaluation of weakly-supervised systems.

7. Acknowledgments

This work was supported in part by the NKTH grant of the Jedlik Ányos R&D Programme 2007 (project codename TUDORKA7 and TEXTREND) of the Hungarian government. The authors wish to thank Orsolya Vincze and Mihály Minkó (the two annotators) and Balázs Tóth (the annotation tool developer) for their devoted efforts.

8. References

[1] B. Adelberg, "NoDoSE - a tool for semi-automatically extracting structured and semistructured data from text documents", *ACM SIGMOD Record Volume 27, Issue 2*, ACM, 1998, pp. 2831-294.

[2] M.E. Califf, and R.J. Mooney, "Relational Learning of Pattern-Match Rules for Information Extraction", University of Texas at Austin, Austin, TX, USA, 1998

[3] E.F. Churchill, "Guest Editors' Introduction: Social Networks and Social Networking", *Internet Computing, Volume 9, Issue 5*, IEEE, 2005, pp. 14-19.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", *Proc. Of the 9th IEEE Intl. Conf. on Tools With Artificial Intelligence (ICTAI'97)*, 1997, pp. 558-567.

[5] D. Freitag, "Information Extraction from HTML: Application of a General Machine Learning Approach", AAAI Press / The MIT Press, Madison, Wisconsin, USA, 1998, pp. 517-523.

[6] G. Hripcsak, and A.S. Rothschild, "Agreement, the F-measure, and reliability in information retrieval", *Journal of the American Medical Informatics Association, Volume 12*, 2005, pp. 296-298.

[7] M. Kayed, and K.F. Shaalan, "A Survey of Web Information Extraction Systems", *IEEE Transactions on Knowledge and Data Engineering, Volume 18*, IEEE Educational Activities Department, 2006, pp. 1411-1428.

[8] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, "Analysing Social Networks Within Bibliographical Data", *Book Series, Lecture Notes in Computer Science, Volume Volume 4080/2006*, Springer Berlin / Heidelberg, 2006, pp. 234-243.

[9] R. Kosala, and H. Blockeel, "Web mining research: a survey", *ACM SIGKDD Explorations Newsletter Archive, Volume 2, Issue 1*, ACM, 2000, pp. 1-15.

[10] N. Kushmerick, "Wrapper induction: Efficiency and expressiveness", *Artif. Intell. Volume 118*, 2000, pp. 15-68.

[11] B. Liu, and K. Chen-Chuan-Chang, "Editorial: special issue on web content mining", *ACM SIGKDD Explorations Newsletter, Volume 6, Issue 2*, ACM, 2004, pp. 1-4.

[12] Y. Matsuoa, J. Morib, M. Hamasakia, T. Nishimuraa, H. Takedab, K. Hasidaa, and M. Ishizukab, "POLYPHONET: An advanced social network extraction system from the Web", *Journal of Web Semantics, Volume 5, Number 4*, 2007, pp. 262-278.

[13] B. Mobasher, "Web Usage Mining", *Encyclopedia of Data Warehousing and Mining*, John Wang (eds.), 2006

[14] E.C.M. Noyons, and A.F.J. van Raan, "Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research", *Journal of the American Society for Information Science archive, Volume 49, Issue 1, Special issue on science and technology indicators*, John Wiley & Sons, Inc., New York, NY, USA, 1998, pp. 68-81.

[15] P.J. Planet, "Tree disagreement: measuring and testing incongruence in phylogenies", *Journal of Biomedical Informatics archive, Volume 39, Issue 1, Special issue: Phylogenetic inferencing: Beyond biology*, Elsevier Science, San Diego, USA, 2006, pp. 86-102.

[16] A.M. Popescu, "Information Extraction from Unstructured Web Text", University of Washington, 2007, pp. 1-152.

[17] C.J. van Rijsbergen, "Information Retrieval", Butterworths, London, 1979.

[18] J. Srivastava, R. Cooley, M. Deshpnde, and P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations*, 2000, pp. 12-23.