An Approach to Content Extraction from Scientific Articles using Case-Based Reasoning

Rajendra Prasath and Pinar Ozturk

Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Slands vei 9, 7491 Trondheim, Norway
drrprasath@gmail.com; pinar@idi.ntnu.no

Abstract. In this paper, we present an efficient approach for content extraction of scientific papers from web pages. The approach uses an artificial intelligence method, Case-Based Reasoning(CBR), that relies on the idea that similar problems have similar solutions and hence reuses past experiences to solve new problems or tasks. The key task of content extraction is the classification of HTML tag sequences where the sequences representing navigation links, advertisements and, other non-informative content are not of interest when the goal is to extract scientific contributions. Our method learns from each experience with the tag sequence classification episode and stores these in the case base. When a new tag sequence needs to be classified, the system checks its case base to see whether a similar tag was experienced before in order to reuse it for content extraction. If the tag sequence is completely new, then it uses the proposed algorithm that relies on two assumptions related to the distribution of various tag sequences occurring in the page, and the similarity of the tag sequences with respect to their structure in terms of levels of the tags. Experimental results show that the proposed approach efficiently extracts content information from scientific articles.

Keywords: Literature Based Knowledge Discovery, Information Extraction, Similar Pattern Mining, Case Based Reasoning

1 Introduction

Humans manage not to get distracted by advertisements, navigational menus, recommended articles, etc when they read a web page because they quickly learn where each of these noisy portions of a page is usually located and how does each portion look like. The ability of distinguishing the actual content from the rest of the page is necessary for computers as well because the noisy text may adversely affect the search and text mining results. We are specifically interested in text mining of scientific papers for the purpose of knowledge discovery. Hence, we would like the computer to process only the content of the scientific articles crawled from scientific publishers. This would have been an easy task if all journal publishers had used the same structure where each type of noisy element was

always located at the same position on a page. However, the fact is that each publisher do it differently and the same publisher may use a different layout for each journal they are publishing. In addition, each publisher often changes the layout they are using over time. Hence, at any moment there are a large number of layouts for the scientific papers on the Web. This makes it necessary for the computer to learn from its experiences and quickly recognize a page layout if it was deciphered earlier.

Several web page information extraction techniques have been introduced to automatically extract the main content using various hybrid approaches that apply segmentation methods based on heuristics or visual features to identify the main content of the webpage. Finding the features that are more salient for recognizing the main content is a challenging task. This problem is much complicated when we attempt to extract coherent content from scientific research articles. In this paper, we present an approach for learning to extract the main content using case based reasoning, an artificial intelligence method for problem solving, that applies incremental learning and reuses its knowledge about specific patterns that it has observed earlier for the efficient extraction of the main content from scientific articles.

This paper is organized as follows: After presenting a brief review of the related work in Section 2, we describe the objectives of the proposed work in Section 3. Then in Section 4, we briefly present the artificial intelligence method called Case-Based Reasoning. In Section 5, the proposed approach is organised in three subtasks: *corpus acquisition*, *patterns extraction*, and *content extraction*. Section 6 describes our experimental results. Finally Section 7 concludes the paper with a brief discussion on future work.

2 Related Work

Web content extraction is very well investigated in the literature [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Many of these approaches apply techniques based on certain heuristics, machine learning or site specific solutions like rule based content extraction, DOM tree parsing, Text graph or Link Graph or vision based models, or NLP features like $N-$grams or shallow text features like number of tokens, average sentence length and so on.

Cai *et* al. [1] presented a compilation of various approaches to handle web content and approaches to perform web mining related tasks. Debnath *et* al. [2] presented an approach for the automatic identification of informative sections of webpages. This approach segments the given web document into several constituent web page blocks and then applies several algorithms to identify the text blocks of the primary content section and filters out the non-informative content.

Finding the content portion of a web page is not straight forward across multiple websites. Many solutions to this problem involve customized rules or scripts to extract the content. Writing separate extraction routines for each website is also time-consuming and many times fail when the web page

layout changes over time. Gibson *et* al. [3] presented an approach that models the content identification problem as a sequence labelling problem using a Conditional Random Field sequence labelling model.

Chakrabarti *et* al. [4] presented an approach for segmenting a webpage into visually and semantically cohesive parts based on weighted graph technique. The weights between two nodes in the Document Object Model (DOM) tree should be placed together or apart in the segmentation. This framework learns the weights from manually labelled data and helps in the segmentation process.

Kohlschütter *et* al. [6] proposed a boilerplate detection algorithm using shallow text features. In this work, non-informative text portions called *boilerplates* have been detected by using a small set of shallow text features. These shallow text features help in classifying the individual text elements in a webpage. Boilerplate creation process is guided by a stochastic model that uses features like: *average word length*, *average sentence length*, *text density*, *link density*, *number of words* and combination of local features.

Sleiman and Corchuelo [9] presented a comprehensive survey of content extractors from web documents. They also proposed an unsupervised web data extractor using Trinary trees. In this approach, two or more web documents from same source are taken to learn a regular expression which is then used to extract data from similar documents. This works on the hypothesis that the input source documents share some common patterns. In another work[13], the same authors proposed an unsupervised information extractor that finds and removes shared token sequences amongst these web documents until finding the informative content.

Yao and Zuo [14] presented a classification based approach to perform webpage content extraction. In this work, a set of relevant features is selected for each text block in the HTML document and then using a Support Vector Machine (SVM) classifier, each text block is classified as either content block or non-content block. Most recently, Wu *et* al. [12] formulated the content identification problem as a DOM tree node selection problem. Using multiple features from DOM node properties, a machine learning model is trained and a set of candidate nodes is selected based on the learning model. The authors observed that the actual content is found in a spatially continuous block. They developed a grouping approach to filter out noisy content and pick the missing data for the candidate nodes.

3 Objectives

With the rapid growth of various research activities and scientific document publishers, the volume of scientific articles keep increasing very fast. Scientist need text mining and search systems to support their knowledge discovery endeavour. The users often search for a topic, a particular researcher/author, combination of topic and year, etc. while in text mining specific entities (e.g., temperature, pH, phytoplankton) or topics (e.g., impacts of increasing temperature on phytoplankton growth) may be focussed on. The input to a text

mining system is the content of the scientific papers on the Web. Hence we aim to develop a content extraction approach that is effective and domain independent.

Figure 1 illustrates a web page of a scientific paper consisting of various blocks where same type of blocks are marked with the same number. Among these blocks, some are just noisy blocks such as navigational links (marked with number 1), banners, personalized menus and advertisements. We need to identify the blocks of interest that cover various parts of a scientific article (marked with number 7 in Figure 1) such as author details, affiliation details, abstract, keywords, headers, sections, subsections, figures, tables, references, contact details, acknowledgements, and so on. To distinguish these blocks from the noisy blocks, we refer the former as *valid* or *informative* blocks of the page. Our goal is to differentiate between blocks with informative content from the ones with non-informative content, and only extract the content of the informative parts of a web page. Figure 1 illustrates how the non-informative or non-interesting parts of the page are spread over the page. A Web page is represented in HTML format which can be translated into a representation consisting of a set of HTML tag sequences where only some correspond to the valid content. The valid content identification can be casted as recognition of valid tag sequences. Tag sequence recognition, as seen in Section 2 is a challenging and costly process based on some regularities and heuristics pertinent to web page layouts.

Our objective is to develop a content extraction system that can efficiently identify valid tag sequences in the web page through acquiring a new experience each time it encounters a tag sequence the first time, and reusing this knowledge when a new tag sequence to be classified (as informative or non-informative) was seen earlier.

4 Case-Based Reasoning

Before describing the proposed approach to content extraction from scientific articles, we give a brief description of the case-based reasoning method which is the backbone of our approach.

Case-Based Reasoning is problem solving paradigm that solves new problems based on the solutions of similar past problems [15, 16, 17]. The past cases are stored in a case base where each *case* is a contextualized piece of experience and has a problem description part which can be represented either in vector representations, structured representation or in text representations[17], and a solution part. CBR consists of four important steps: *Retrieve*, *Revise*, *Reuse* and *Retain* (see Figure 2). When used for a classification task, the problem description will consist of the features describing the problem while the solution will be the class that it belongs to while in a planning task, the solution will be a sequence of actions to take. The principle idea for all task is that similar problems will have similar solutions. Hence, if a new problem description matches with an already solved problem existing in the case base it is retrieved to be re-used for solving the new problem. CBR allows incremental learning; when the current

Fig. 1. Scientific article crawled from Nature Publishers. Blocks marked with X are not part of the main article content and need to be filtered out. Only those marked with number 7 are "valid/informative" for our purpose.

problem to be solved does not already exist in the case base, it is retained in the Case Base after it is solved in order to be reused to solve future similar problems.

We use CBR for gathering informative tag sequence experiences and using these in order to increase the efficiency of the system in classifying tag sequences in a new web page.

5 Proposed Approach

The overall approach consists of three main components, as shown in Figure 3. The first is the *corpus acquisition* component which converts the HTML

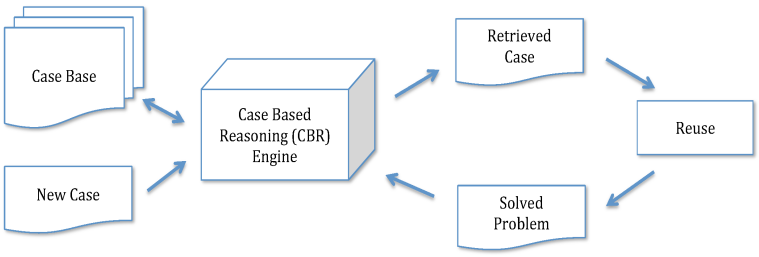

Fig. 2. Case Based Reasoning Paradigm

representation of a web page that includes a scientific text into a representation consisting of a set of HTML-tag sequences where each tag sequence corresponds to a block (labelled with a number in Figure 1) in the page. Figure 1 illustrates different blocks of a web page of a scientific paper. The next component, *pattern extraction*, classifies a tag sequence as either informative/interesting or non-informative. The last component is *content extraction* which extracts the content of the blocks in the web page represented by valid patterns. The key issue is to identify tag-sequences in HTML representation that are corresponding to the valid content, e.g., sections of the scientific article. For the knowledge discovery task, we may also be interested in some information about the article, such as authors and the journal it is published in (marked with uppermost 7 in Figure 1).

Our approach includes an algorithm that classifies a tag sequence as representing valid or noisy data, through an elaborative comparison of all tag-sequences in the concerned webpage (explained in detail in Section 5.2.2. An important attribute of the proposed approach is that it can take advantage of its earlier experiences about tag sequence classification. The tag sequence classification in Section 5.2.2 is used only when the new tag to be classified is not seen before, that is, a case similar to this one does not exist in the case base of the system.

The overall approach is represented as a pseudocode in Algorithm 1 while Figure 3 gives a more detailed picture of the approach. In the rest of Section 5, we describe three main components and their sub-components in detail.

5.1 Corpus Acquisition

5.1.1 Crawling We have identified the set of seed urls for the articles of the journals that are considered under three different categories in the specified domain by domain experts: *Level - 2: high impact journal*, *Level - 1: low important journal* and *Level - 0: other journals*. The choice of the categories may vary across researchers in the same domain and the impact of the journal is also selected based on their scientific impact factor and the citation of the articles published in that journal. So it is a co-related factor associated with the choice of the publication channel of the domain experts in the specific domain. We have used the collected seed urls and performed crawling of scientific articles. In

Algorithm 1 The proposed approach

Input: A collection of scientific articles, each in HTML format

Description:

1: **Preprocess** the input HTML document into a well structured HTML document
2: **Parse** the HTML document into a set of tag-sequences
3: **for** each tag sequence of the document **do**
4: **Match** the tag sequence with the cases in the case base
5: **if** a similar case (i.e, a tag sequence / pattern) is retrieved **then**
6: Perform content extraction with the retrieved pattern
7: **else**
8: **Identify the candidate tag sequence / pattern** using count- and level-assumptions and retain in the case base
9: **Extract** the content using the identified pattern
10: **end if**
11: **end for**

Output:

a) A case base of tag patterns;
b) Extracted blocks of the main content of scientific articles.

this experiment, we have considered only a few publishers popular in publishing journals of high impact factors. After crawling the scientific articles, we have stored the original content of the articles in HTML format. We further use these articles to perform the extraction of the main content for exploratory purposes.

5.1.2 Preprocessing The scientific articles are crawled in HTML format from various publishers. An HTML document in the crawled data may have noisy contents like advertisements, banners, recommended articles, navigational links, copyright information and so on. We perform the metadata extraction of the article in the form of an attribute-value pairs. Additionally, we use the pattern extractor that explores the structure of the main article content, filters out the patterns with class references of similar blocks of the main text content of the scientific article. The text content of such similar patterns can represent coherent blocks (or sections) of a scientific article. We have adapted the preprocessing steps similar to the one given in [18].

In Section. 5.3.1, we explain the detailed procedure to segment an article (in HTML format) into coherent blocks.

5.1.3 Parsing We apply top down parsing by traversing over the HTML tags (nodes) in the well structured HTML document. At each node, we populate the tag sequence of the underlying subtree and we repeat this process until we process all open tags till the end of the web document. Each of these tag sequences represent a specific part of the layout. So the output of the parsing is a set of tag sequences.

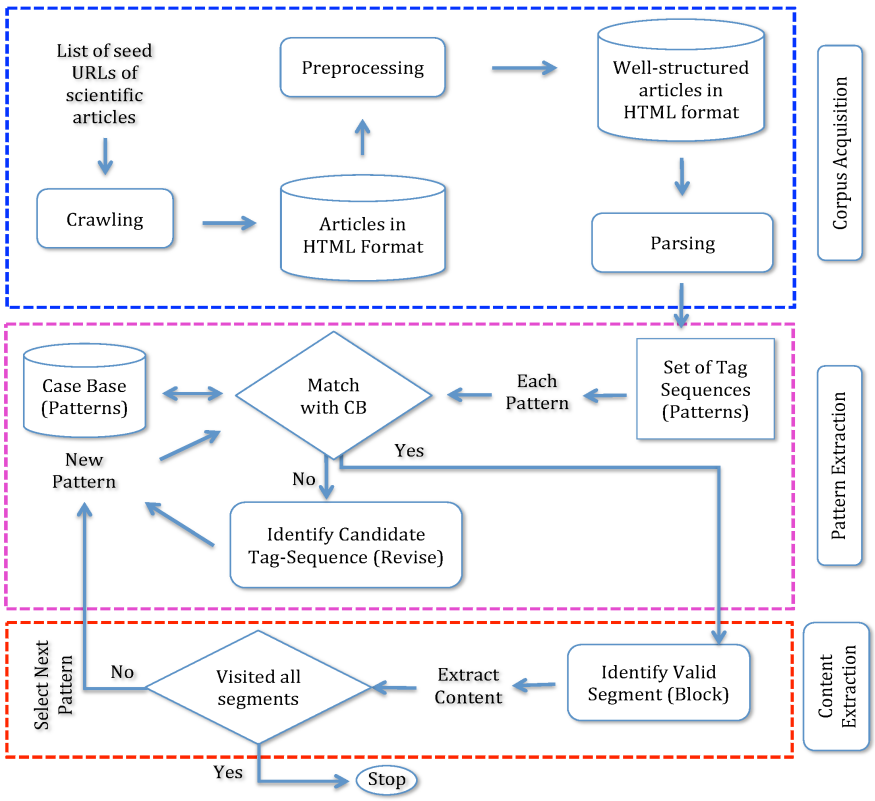

Fig. 3. The proposed CBR-based approach to content extraction.

5.2 Pattern Extraction

5.2.1 CBR Retrieval When a new tag sequence in the document is to be classified as a valid or invalid pattern, the system first tries to retrieve a case with a similar pattern from the case base. Since we preserve the order of tags in the sequence, the similarity measure applied here is based on substring matching and gives higher preference to the matches that occur at higher levels of tags. If the system finds a similar case, then the current tag sequence is valid indicating that the corresponding content should be extracted. If there is no similar case in the case Base, then the system resorts to the classification algorithm explained in Section 5.2.2. If the tag sequence is classified as valid by this algorithm, then it is learnt and retained in the case base.

5.2.2 Identifying Candidate Tag Sequences Our approach is based on two important assumptions that we call as *count* assumption and *level* assumption. The *count assumption* expresses that sections of a scientific article are assumed to follow a similar tag sequence. This means that this type of tag

sequence occurs as many times as the number of sections in the article. This in turn, means that the tag sequence with the most occurrence on a web page of the scientific article is highly probable to belong to the scientific text. The *level assumption* suggests that for an article with several sections, even when the whole tag sequence of each section does not fully matches with that of other sections, a large portion of tag sequences of all sections may be overlapping at different levels that correspond to different subsections. Tag sequences of two sections that overlap in higher levels are considered to be more similar. The more such higher levels they overlap, the more similar they would be. Putting these two assumptions together, tag sequences that are similar according to the level assumption that occurs most often in a web page are good candidates to represent a section of a scientific text. These candidate patterns are used to populate the case base which was empty at the initial state.

5.3 Content Extraction

5.3.1 Segmenting Scientific Article into Blocks

For each valid tag sequence in the document, we scan through the document and identify all matching blocks. Each block is segmented according to the tag sequence.

5.3.2 Extracting Text from Blocks Step 1: We extract the meta tags by matching META attributes and values of the HTML document. We collect most of the available metadata of articles from the HEAD tag.
Step 2: In this step, we focus on the content part of the scientific article. We match the tag sequences to the corresponding blocks of the scientific articles. We extract the text from matching blocks.

6 Experimental Results

6.1 Corpus

We have crawled scientific articles published by various publishers and the details are listed in Table 1. In this experiment, we have crawled both open access articles and articles with restricted access option (subscription based). The articles collected from journals: *Climate Change*, *Ecosystems* and *The ISME Journal* are of open access and the articles collected from journals: *Ecology Letters*, *Nature Geoscience* and *Nature Communications* are of restricted access.

Scientific research articles in the areas of Marine Science, Climate Science and Environmental Science are the outcome of research experiments and observations describing the underlying complex theories and models. Lots of variables like chemical compositions, metals, living organisms and other species are involved in such environmental models. To explore the facts behind these variables and find out how the variables are quantified in the presence of an associated variable or an event in the marine food chain. These journals are suggested by the

Publisher	Journal Name	# Articles
Springer	Climate Change	473
Springer	Ecosystems	73
Macmillan Publishers	The ISME Journal	113
Wiley	Ecology Letters	1650
Macmillan Publishers	Nature Geoscience	225
Macmillan Publishers	Nature Communications	2,741

Table 1. Data used in our experiments

domain experts. This data is the part of the ongoing research work on the literature based knowledge discovery tasks and is a part of the corpus used in the OCEAN-CERTAIN[1] project.

6.2 Evaluation Measures

In this section, we present the evaluation measure of the proposed approach. We used the following measures to quantify the goodness of the extracted blocks: *Purity* and *Accuracy*. Let us define these measures now.

Purity is used to quantify the number of noise free blocks that are extracted from the given input document. We define *Purity* as the ratio between the number of blocks without noisy content and the total number of blocks extracted.

Accuracy is used to quantify the number of content blocks that are correctly extracted from the given input document. We define *accuracy* as the ratio between the number of blocks correctly extracted and the total number of blocks in the input document.

6.3 Discussion

We have presented our preliminary results in this work. We have considered top 20 articles in each journal and manually evaluated the efficiency of the proposed information extraction approach. The results are tabulated in Table. 2.

Journal Name	Purity #	Accuracy
Climate Change	91.2 %	95.43 %
Ecosystems	93.57 %	88.52 %
The ISME Journal	90.72 %	91.83 %
Ecology Letters	85.27 %	89.39 %
Nature Geoscience	94.26 %	93.41 %
Nature Communications	96.37 %	95.43 %

Table 2. Purity and Accuracy scores of the selected journals (averaged over 20 documents)

[1] OCEAN CERTAIN - A project funded by European Union and lead by NTNU with 11 partners from 8 European countries and Chile and Australia. `http://cordis.europa.eu/project/rcn/110540_en.html`

Since Nature articles are well organized into coherent sections, each section heading and subheadings are explicitly marked up. It also helps to filter out figure captions along with the table related data from the main content. In journals published by other publishers, the wrapper has to identify and extract this information with more efforts. Additionally, we observed majority of the articles consists of citation and copyright blocks along with the main content. It also reduces the accuracy of the proposed approach.

7 Conclusion

In this work, we have proposed an approach that performs learning to extract vital information from scientific articles based on a Case-Based Reasoning approach. In a web page of a scientific paper, there will be more number of article related blocks than other, noisy parts. Since tag sequences representing the segments of the scientific text will be similar, the tag sequence with highest number of occurrence in the page is a good candidate to represent the content of our interest. Here we attempted to use this heuristic without relying on any linguistic or semantic expertise to extract the right content from different publishers. Experimental results carried out on the subset of the scientific articles collection show that the proposed approach effectively extracts the relevant information that is useful to do knowledge discovery. Further we need to recognize patterns that belong to the specific parts of a scientific article, that is, whether the pattern belongs to author(s) name, title or the main body of the scientific content. To achieve this, we plan to extend the CBR part of the system and add the label of the tag sequence that specifies the part of the scientific article.

Acknowledgment

Authors gratefully acknowledge the support of European Commission through OCEAN-CERTAIN under the grant no. 603773.

References

[1] Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Extracting content structure for web pages based on visual representation. In: Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications. APWeb'03, Berlin, Heidelberg, Springer-Verlag (2003) 406–417

[2] Debnath, S., Mitra, P., Pal, N., Giles, C.L.: Automatic identification of informative sections of web pages. IEEE Trans. on Knowl. and Data Eng. **17**(9) (September 2005) 1233–1246

[3] Gibson, J., Wellner, B., Lubar, S.: Adaptive web-page content identification. In: Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management. WIDM '07, New York, NY, USA, ACM (2007) 105–112

[4] Chakrabarti, D., Kumar, R., Punera, K.: A graph-theoretic approach to webpage segmentation. In: Proceedings of the 17th International Conference on World Wide Web. WWW '08, New York, NY, USA, ACM (2008) 377–386
[5] Pasternack, J., Roth, D.: Extracting article text from the web with maximum subsequence segmentation. In: Proceedings of the 18th International Conference on World Wide Web. WWW '09, New York, NY, USA, ACM (2009) 971–980
[6] Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. WSDM '10, New York, NY, USA, ACM (2010) 441–450
[7] Sleiman, H.A., Corchuelo, R.: An unsupervised technique to extract information from semi-structured web pages. In Wang, X.S., Cruz, I.F., Delis, A., Huang, G., eds.: Web Information Systems Engineering - WISE 2012 - 13th International Conference, Paphos, Cyprus, November 28-30, 2012. Proceedings. Volume 7651 of Lecture Notes in Computer Science., Springer (2012) 631–637
[8] Sleiman, H.A., Hernández, I.: A framework for populating ontological models from semi-structured web documents. In Atzeni, P., Cheung, D.W., Ram, S., eds.: Conceptual Modeling - 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings. Volume 7532 of Lecture Notes in Computer Science., Springer (2012) 578–583
[9] Sleiman, H.A., Corchuelo, R.: A survey on region extractors from web documents. IEEE Trans. Knowl. Data Eng. **25**(9) (2013) 1960–1981
[10] Sleiman, H.A., Corchuelo, R.: Trinity: On using trinary trees for unsupervised web data extraction. IEEE Trans. Knowl. Data Eng. **26**(6) (2014) 1544–1556
[11] Crescenzi, V., Merialdo, P., Qiu, D.: Crowdsourcing large scale wrapper inference. Distributed and Parallel Databases **33**(1) (2015) 95–122
[12] Wu, S., Liu, J., Fan, J.: Automatic web content extraction by combination of learning and grouping. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, New York, NY, USA, ACM (2015) 1264–1274
[13] Sleiman, H.A., Corchuelo, R.: TEX: an efficient and effective unsupervised web information extractor. Knowl.-Based Syst. **39** (2013) 109–123
[14] Yao, J., Zuo, X.: A machine learning approach to webpage content exraction. `http://cs229.stanford.edu/proj2013/YaoZuo-AMachineLearningApproachToWebpageContentExtraction.pdf` (2013) Online; Last accessed 23 February 2016.
[15] Kolodner, J.: Case-based Reasoning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
[16] Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Commun. **7**(1) (March 1994) 39–59
[17] Bergmann, R., Kolodner, J., Plaza, E.: Representation in case-based reasoning. Knowl. Eng. Rev. **20**(3) (September 2005) 209–213
[18] Prasath, R.R., Öztürk, P.: Similarity assessment through blocking and affordance assignment in textual cbr. In: Proc. of the Reasoning from Experiences on the Web. WebCBR 2010, Alessandria, Italy, Web CBR (2010) 51–160