Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 27–34
DOI: 10.15439/2014F492
ACSIS, Vol. 2

Adaptive Learning for Improving Semantic Tagging of Scientific Articles

Andrzej Janusz
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
janusza@mimuw.edu.pl

Sebastian Stawicki
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
stawicki@mimuw.edu.pl

Hung Son Nguyen
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
son@mimuw.edu.pl

***Abstract*—In this paper we consider a problem of automatic labeling of textual data with concepts explicitly defined in an external knowledge base. We describe our tagging system and we also present a framework for adaptive learning of associations between terms or phrases from the texts and the concepts. Those associations are then utilized by our semantic interpreter, which is based on the Explicit Semantic Analysis (ESA) method, in order to label scientific articles indexed by our SONCA platform. Apart from the description of the learning algorithm, we show a few practical application examples of our system, in which it was used for tagging scientific articles with headings from the MeSH ontology, categories from ACM Computing Classification System and from OECD Fields of Science and Technology Classification.**

***Keywords*—semantic indexing; Explicit Semantic Analysis; Adaptive Semantic Analysis; multi-label tagging; adaptive learning;**

I. Introduction

THE MAIN idea of a keyword search is to look for texts (documents) that contain one or more words specified by a user. Then, using a dedicated ranking algorithm, relevance of the matching documents to the user query is predicted and the results are served as an ordered list [1]. In contrast, semantic search engines try to improve the search accuracy by understanding both, the user's information need and the contextual meaning of texts, which are then intelligently associated [2].

From the data processing point of view, the semantic search engine may be divided into three main components: semantic text representation module, interpretation and representation of a user query, and an intelligent matching algorithm [3]. The scope of the first two modules may be categorized as a semantic data representation [4]. In opposite to the keyword search, the semantic data representation, and thus the semantic indexes, cannot be calculated once and then utilized by intelligent matching algorithms. The text representation, as well as a query interpretation should be assessed with respect to the type of the users' group, a context of the words in the query and many others factors [5].

This work is partially supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information" and by Polish National Science Centre (NCN) grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215.

The better part of current search engines is based on a combination of a keyword search and sophisticated document ranking methods [1]. Only some of them process search queries, analyzing both, a query and documents' content with respect to their meaning, and return the semantically relevant search results [2]. However, even this approach becomes insufficient. The process of information retrieval needs to be made intelligently in order to help users in finding relevant information. The key role in this process is the recognition of the users' information needs and collecting feedback about the search effectiveness. The gathered information should be utilized to improve search algorithms and forge better responses to user requirements. Those challenges are in the scope of studies on adaptive search engines which interact with experts (users) and operate in a semantic representation space [6].

The SONCA (Search based on ONtologies and Compound Analytics) platform [7] is developed at the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw. It is a part of SYNAT project focusing on development of Interdisciplinary System for Interactive Scientific and Scientific-Technical Information (www.synat.pl). SONCA aims at extending the functionality of search engines by more efficient search of relevant documents, intelligent extraction and synthesis of information, as well as a more advanced interaction between users and knowledge sources.

Within the SYNAT project, some successful methods for the semantic text representation and indexing have already been developed [4], [8]. In this paper we discuss an adaptive learning model of terms-to-concepts associations which can be treated as an extension of the Explicit Semantic Analysis (ESA) method [9], [10]. By an analogy, we call it Adaptive Semantic Analysis (ASA). The main purpose of this model is to adjust the links between words and well-defined concepts. Those links are automatically derived from natural language definitions of the concepts which are stored in an external knowledge base. The associations are then used for labeling and indexing scientific articles. The definitions of the concepts can be extracted from different knowledge sources such as domain ontologies. In our experiments we show how the model can be constructed using descriptions of the concepts in a natural language and how it can be improved by using feedback from domain experts. We also show how to deal with a lack of concept descriptions in a case when there is available a sufficient number of training examples of labeled articles. Finally,

978-83-60810-58-3/$25.00

we present our most recent developments and improvements to the model, which are related to a problem of deciding how many concepts should be associated with a given document. As case studies we use a task of tagging biomedical articles from the PubMed repository with concepts from the MeSH ontology [11], a task of labeling abstracts of computer-science-related documents with terms from ACM Computing Classification System [12] and a problem of assigning the OECD Fields of Science and Technology Classification categories to articles from the Infona system (www.infona.pl).

II. Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) proposed in [9] is a method for automatic tagging of textual data with Wikipedia concepts. It utilizes natural language texts of Wiki articles as textual representations of the corresponding concepts. It is assumed that the Wiki articles contain definitions of the concepts and describe their semantic. Those representations are regarded as a regular collection of texts and are matched against documents to find the best associations [10].

In ESA, the semantic relatedness between concepts and documents is computed two-fold. First, after the initial processing (tokenization, stemming, stop words removal), the corpus and the concept definitions are converted to the *bag-of-words* representation. Each of the distinct terms in the documents is given a weight expressing a strength of its association to the text. Assume that after the initial processing of a corpus consisting of M documents, $D = \{D_1, \ldots, D_M\}$, there have been identified N distinct terms (e.g. words, stems, n-grams) $t_1, \ldots, t_N$. Any text D_i in the corpus D can be represented by a vector $W_i = \langle w_{1,i}, \ldots, w_{N,i} \rangle \in \mathbb{R}_+^N$, where each coordinate $w_{j,i}$ expresses a value of some relatedness measure for j-th term in vocabulary (t_j), relative to this document. The most common measure used to calculate $w_{j,i}$ is the *tf-idf* (term frequency-inverse document frequency) index [1], defined as:

$$w_{j,i} = tf_{i,j} * idf_j = \frac{n_{i,j}}{\sum_{k=1}^{N} n_{i,k}} \log\left(\frac{M}{|\{i : n_{i,j} \neq 0\}|}\right), \quad (1)$$

where $n_{i,j}$ is the number of occurrences of the term t_j in the document D_i.

In the second step, the bag-of-words representation of the concept definitions is transformed into an inverted index that maps the terms $t_1, \ldots, t_N$ into lists of K concepts $C_1, \ldots, C_K$, described in an external knowledge source. The inverted index can be used as a semantic interpreter. Given a document from the corpus D, we may iterate over terms from the text, retrieve the corresponding entries from the inverted index and merge them into a vector of concept weights that represents the analyzed document.

Let $W_i = \langle w_{1,i}, \ldots, w_{j,i}, \ldots, w_{N,i} \rangle$ be a bag-of-words representation of an input document D_i, where $w_{j,i}$ is the tf-idf index of t_j defined by the formula (1). We can analogically quantify the association between the term t_j and the k-th concept C_k by computing the *bag-of-words* representations of the concept descriptions. Those associations constitutes the inverted index. Let $inv_{j,k}$ be the inverted index entry for the term t_j and the concept C_k. For convenience, all the weights $inv_{j,k}$ can be arranged in a sparse matrix structure with N rows and K columns, denoted by INV, such that $INV[j,k] = inv_{j,k}$ for any pair (j,k), such that $j = 1, \ldots, N$ and $k = 1, \ldots, K$. The new vector representation of D_i will be denoted by $V_i = \langle v_{1,i}, \ldots, v_{K,i} \rangle$ where:

$$v_{k,i} = \sum_{j : t_j \in D_i} w_{j,i} * inv_{j,k}. \quad (2)$$

In other words, the above equation expresses a standard dot product of the k-th column of the matrix INV and the vector W_i. This new representation will be called a *bag-of-concepts* of a document D_i.

For practical reasons it may also be useful to represent documents only by the most relevant concepts. In such a case, the association weights can be used to rank the concepts and to select only the top concepts from the ranked list. One can also apply some more sophisticated methods that involve utilization of internal relations in the knowledge base (e.g. for semantic clustering of concepts and assigning only the most representative ones to the documents).

The original purpose of Explicit Semantic Analysis was to provide means for computing semantic relatedness between texts. However, an intermediate result – weighted assignments of concepts to documents (induced by the term-concept weight matrix) may be naturally utilized in document retrieval as a semantic index [3], [5]. Although originally ESA was meant to utilize the Wikipedia articles as the external knowledge source, it seems reasonable that for specialized tasks, such as indexing articles from a specific branch of science, it is better to use concepts described in dedicated knowledge bases or ontologies. A user (an expert) may query a document retrieval engine for documents matching a given ontology concept. If the concepts are already assigned to documents, this problem is conceptually trivial. However such a situation is relatively rare, since the employment of experts who could manually labeled documents from a huge repository is expensive. On the other hand, the utilization of an automatic tagging method, such as ESA, allows to infer a labeling of previously untagged documents or at least it can support the experts in that task.

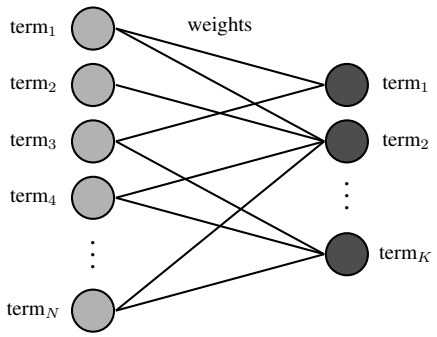

Fig. 1. A schema of the inverted index utilized by ESA.

III. The Adaptive Learning Algorithm

During our research on the automatic tagging methods we noticed that the structure of the inverted index used by ESA resembles a structure of an artificial neural network [13]. This network consists of a single layer of perceptrons (neurons)

that correspond to the different concepts and their inputs correspond to the distinct terms from the concept definitions. However, unlike in a classical neural network, in the case of ESA the inputs are not connected with every neuron. In fact, the net of the connections is rather sparse, since only a small fraction of terms appears in a single description of a concept. Each connection has a non-negative weight from the inverted index, which quantifies the association between the term and the concept. Such a schema of the inverted index structure is depicted in Figure 1.

In the classical neural networks, the activation of each neuron is determined by computing a value of, so called, an activation function. This function takes as its argument a weighted sum of input values and returns a real-valued output. In ESA the activation function corresponds to the identity function. Nevertheless, it is possible to use other types of functions, including those which are typically used in perceptrons (e.g. sigmoid, hyperbolic tangent [13]), in order to scale the association values into a desired range. We can also easily modify the network structure from Figure 1 by adding an additional input, connected to each of the neurons. This input will be treated as an activation threshold. We will assign a concept C_k to a document D_i only if its association $v_{k,i}$ exceeds the corresponding activation threshold a_k, $k = 1, \ldots, K$.

Since the model resembles a neural network, in our previous research [8] we proposed to use a simple learning algorithm for the adaptation of weights from the inverted index to a feedback regarding the tagging quality, obtained from domain experts. The algorithm was based on a typical perceptron learning schema, namely the error backpropagation approach [14]. It is shown in Figure 2. Here we present an improved version of this algorithm that does not require a prior information regarding a number of concepts that should be assigned to each document. For this purpose, we first need to discuss the types of errors that can be made in predicting a set of labels that should be assigned to a given document.

Let us denote by $esa(D_i, INV, A)$ a set of concepts assigned by ESA to a document D_i, using the inverted index INV and the vector of activation thresholds A. This set consists of those concepts whose associations to D_i exceeded the activation threshold values, i.e., $esa(D_i, INV, A) = \{C_k : v_{k,i} > a_k, k = 1, \ldots, K\}$. We assume that there is available a corpus D of training documents, for which we can get the sets of truly related concepts. Since those sets of reference labels usually have to be obtained from domain experts, we will denote them by $exp(D_i)$.

Knowing the sets $esa(D_i, INV, A)$ and $exp(D_i)$ we can divide their union into three mutually disjoint subsets: $TP_i = esa(D_i, INV, A) \cap exp(D_i)$, $FP_i = esa(D_i, INV, A) \setminus exp(D_i)$ and $FN_i = exp(D_i) \setminus esa(D_i, INV, A)$. They can be interpreted as the sets of *Truly Positive*, *Falsely Positive* and *Falsely Negative* cases in the classical machine learning theory [13]. The set TP_i contains truly relevant concepts which were assigned by ESA. Since we want to maximized its cardinality, in every iteration of the learning algorithm we will increase the weights of the connections between the terms t_j from D_i and the concepts from TP_i. The update will be proportional to the association strength between the terms and D_i, which is quantified by the values of $w_{j,i}$. Analogically, we will increase the weights of the concepts from FN_i and decrease those of the concepts from FP_i. At the same time we will be updating the activation thresholds in order to move the concepts from FN_i into TP_i and to remove the FP_i concepts from the set $esa(D_i, INV, A)$. Details of this procedure are explained by Algorithm 1. We call it Adaptive Semantic Analysis (ASA) by an analogy to the ESA algorithm.

We impose one constraint on the weight refinement procedure. Only the available concept descriptions determine the network structure of the inverted index. During the learning procedure we do not construct any new connections in the network, i.e. we restrict the weights $inv_{j,k}$ equal zero to remain zero for a whole learning process. Moreover, the updates in our algorithm are multiplicative, which guarantees that $inv_{j,k} \geq 0$ for every j and k. This restriction is motivated by an intuition that the original concept descriptions, which are usually provided by domain experts, contain sufficient vocabulary to characterize the concepts, thus they define a good model of the terms-to-concepts relations. Additionally, by tuning a large number of weights it is possible to fall into a trap of over-fitting the inverted index to the reference data. Moreover, the reduced number of connections in the inverted index makes the learning more efficient, since there are needed considerably less updates at every iteration of the ASA algorithm.

In the algorithm, the activation thresholds are tuned along the concept weights. In practice, however, they do not need to be updated in every iteration. In order to speed up the learning process, the line number **(31)** of Algorithm 1 can be executed periodically, with the length of the period controlled by an additional parameter.

IV. Experiments

We tested our multi-label tagging system on three different problems, namely automatic labeling of biomedical articles from the PubMed Central repository with headings from the MeSH ontology, assigning categories from ACM Computing Classification System (ACM CCS) to articles from ACM Digital Library and labeling research papers from the Infona repository [6] with the OECD Fields of Science and Technology Classification (OECD FOS) categories. In all those experiments we followed the same testing methodology. We split the available corpus into a training and a test set. We use the training data for adaptive learning of the associations between terms and concepts with the proposed ASA algorithm (see Section III), and then we verify the performance of our tagging system on the test data. We repeat the whole procedure several times with different divisions of the data and report the average results. As the quality measures we use average values of the F_1-$score$, $Precision$ and $Recall$, obtained for all the test documents by comparing the predicted tags to those which were assigned by experts or authors. This type of evaluation of a tagging quality is popular for the multi-label classification problems [15].

A. Experiment on Biomedical Articles

In our first series of experiments we performed the tests on a corpus from the PubMed Central repository [16], consisting of roughly $38,000$ publicly available articles. As the external knowledge base we used the MeSH ontology [11],

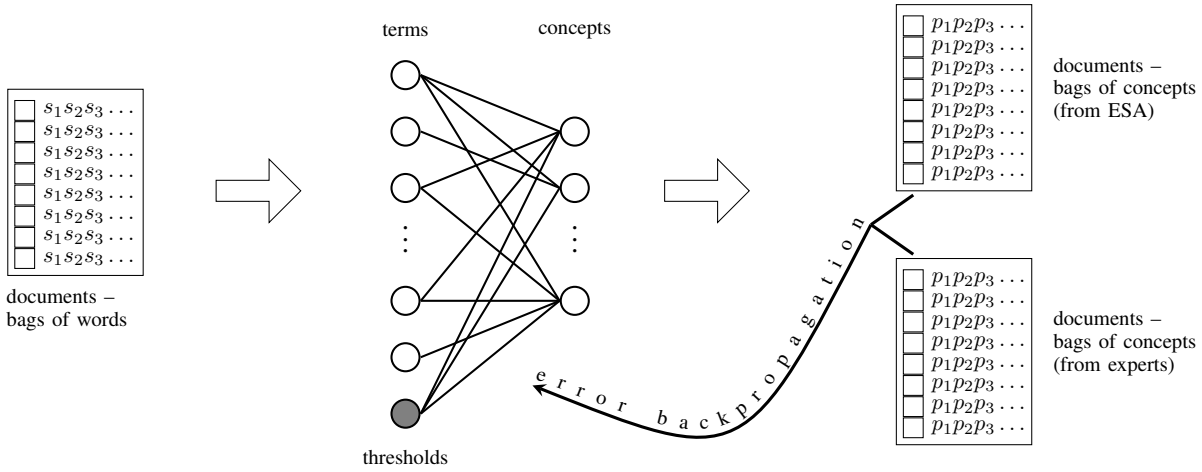

Fig. 2. The learning schema for Adaptive Semantic Analysis.

which is also employed by PubMed to index articles and to facilitate search through its resources. We adapted the ESA method to enable tagging documents from the corpus with the MeSH concepts (also known in MeSH terminology as the headings). In the MeSH ontology, each heading is accompanied by a descriptor data record prepared by domain experts. We composed the final natural language description of the MeSH headings by concatenating the following elements of the corresponding MeSH record: *MESH HEADING* (the name), *MESH SCOPE NOTE* (a short textual description), *ENTRY* (synonyms), *PREVIOUS INDEXING* (previous naming) and *PHARMACOLOGICAL ACTION* (known pharmacological activity). We processed those descriptions using text mining tools in order to determine the initial inverted index structure of our model (i.e. relations between the terms and concepts) and the initial values of the weights. In the experiments we used the edition of MeSH from the year 2012, which contained records on 26.142 main concepts (the headings).

Additionally, for each document in the corpus we obtained sets of MeSH headings assigned by experts from the U.S. National Library of Medicine. The average number of tags assigned to a single article by the NLM experts was ≈ 13.5. We treat those tags as a reference and we utilize them for improving the terms-to-concepts associations with the adaptive learning algorithm described in Section III. We also used those tags for the evaluation purpose. In the tests, we run the adaptation of the inverted index on randomly selected $20,000$ documents and then we use it to tag the remaining part of the corpus. As the starting activation thresholds we use a vector with all coordinates equal 5. This value was chosen using a common sense, based on an observation of a distribution of the concept associations for several exemplary documents. We assess the quality of the tagging by computing the average values of F_1-$score$, $Precision$ and $Recall$ measures. Results of those tests are shown in Figure 3.

The results of the tests turned out to be very promising. On the test data we observed a significant improvement of performance over the regular ESA (the iteration number 0 in the plots) in terms of the computed statistics. For instance, F_1-$score$ value improved by approximately 160% (from ≈ 0.15 to ≈ 0.39). Even a greater improvement was noticed with regards to the values of $Recall$. After the last iteration, its average value exceeded 0.43 while for the regular ESA the average $Recall$ was ≈ 0.16. This however, can be partially explained by the fact that in the initial learning iterations the resulting tagging model usually returned a lower number of labels than the experts.

B. Experiment on Papers from ACM Digital Library

This experiment was conducted on a corpus consisting of publicly available meta-data entries for articles from ACM Digital Library. The corpus contain information on approximately $400,000$ research papers from the field of computer science. The available meta-data included a title, an abstract and in some cases a list of key phrases assigned by authors. We concatenated those information for each document into a single text and we used it to compute its bag-of-words representation. Additionally, the data contained a list of associated ACM CCS categories which also were inputted by the authors. On average, every article was associated with only three out of 1571 possible categories.

The task in this experiment was to label the articles with the ACM CCS categories based on the remaining meta-data. Unlike in the previous experiment, however, this time we did not possess any additional knowledge base with natural language descriptions of the concepts. To deal with this problem we had to slightly modify the procedure of our experiment. After the initial division of the data into the training and test sets (in proportion of 1:1), we divided the training data into two separate sets. For each of the ACM CCS categories we concatenated into a single text the meta-data of all articles that were labeled with this category by the authors. In this way we obtained the textual representation of the categories that could be used for the computation of the initial term-to-concepts associations for our tagging system.

In the second step, we used those associations as a starting point for the ASA algorithm. We performed the adaptive learning of the associations on the remaining part of the training data. We initiated the learning process with the activation thresholds set to 0.30 for all the categories. The starting value of this parameter was much lower than in the experiments

Algorithm 1: Computation of a new inverted index matrix INV^{l+1} and activation thresholds A^{l+1} in the l-th iteration of the adaptive learning algorithm (ASA).

Input: A corpus $D = \{D_i : i \in 1, \ldots, M\}$; INV^l; activation thresholds $A^l = \langle a_1, \ldots, a_K \rangle$;
Output: An updated matrix INV^{l+1}; a vector A^{l+1};

begin
  Initiate ΔINV and CU as empty $N \times K$ matrices;
  Initiate ΔA as a zero vector of length K;
  for $i = 1$ **to** M **do**
    $TP_i = esa(D_i, INV^l, A^l) \cap exp(D_i)$;
    $FP_i = esa(D_i, INV^l, A^l) \setminus exp(D_i)$;
    $FN_i = exp(D_i) \setminus esa(D_i, INV^l, A^l)$;
    foreach $C_k \in esa(D_i, INV^l, A^l) \cup exp(D_i)$ **do**
      $tIds = \{j : t_j \in D_i \wedge INV^l[j,k] > 0\}$;
      $wSum = \sum_{j \in tIds} w_{j,i}$;
      if $C_k \in FP_i$ **then**
        foreach $j \in tIds$ **do**
          $\Delta INV[j,k] = \Delta INV[j,k] - INV^l[j,k] * w_{j,i} / wSum$;
          $CU[j,k] = CU[j,k] + 1$;
        $\Delta A[k] = \Delta A[k] + A^l[k] * \left(1 - \frac{|TP_i|}{|TP_i \cup FP_i|}\right)$;
      else
        foreach $j \in tIds$ **do**
          $\Delta INV[j,k] = \Delta INV[j,k] + INV^l[j,k] * w_{j,i} / wSum$;
          $CU[j,k] = CU[j,k] + 1$;
        $\Delta A[k] = \Delta A[k] - A^l[k] * \left(1 - \frac{|TP_i|}{|TP_i \cup FN_i|}\right)$;
  foreach (j,k) *such that* $CU[j,k] > 0$ **do**
    $\Delta INV[j,k] = \Delta INV[j,k] / CU[j,k]$;
  $INV^{l+1} = INV^l + \Delta INV$;
  $A^{l+1} = A^l + \Delta A / M$;
  return INV^{l+1} *and* A^{l+1}

on biomedical articles due to a fact that this time we had to operate on significantly shorter texts. We measured the quality of our tagging system by comparing the labels assigned to the test articles with the labels which were given by the authors. The results of those comparisons in the consecutive iterations of the learning algorithm are depicted on Figure 4.

Similarly to the previous experiments, the results clearly show that the learning algorithm significantly improves the quality of the tagging system in comparison to the standard ESA (the iteration number zero on the plots). In the test, the highest F_1-$score$ on the test set was ≈ 0.20. It was obtained in the last iteration (50-th) of the algorithm, which suggests that it would be possible to slightly improve the results by giving the algorithm some more time for learning. There is also a noticeable difference between the results on the training and test sets. The highest training F_1-$score$ exceeded 0.33, which is over 50% higher than the corresponding test score. On one hand this difference may be partially explained by the fact that authors do not follow any strict rules or guidelines when they assign the categories to their papers. It makes the assigned labels very subjective. As a consequence, the prediction of the ACM CCS categories becomes a very difficult task. On the other hand, the differences in the tagging quality for the training and test data may be caused by the way we generated the textual descriptions of the ACM CCS categories. The concatenation of many article abstracts had to result in the inclusion of many highly specialized terms into the descriptions. Such terms often allow to identify individual papers, thus their presence may lead to the over-fitting of the learning algorithm to the training data.

C. Experiment on Data from the Infona System

The last series of experiments was conducted on a corpus obtained from the Infona repository [6]. Infona contains meta-data of over 1.8 million articles from a wide range of science fields. However, in our experiments we were restricted to only a small sample of all the data from this repository, i.e. our corpus contained information from 1000 meta-data entries. Each entry consisted of an article title in English, author names and an English abstract. Additionally, for many entries there were available key phrases assigned by the authors. Similarly as in the previous experiments, for each article we concatenated the available meta-data (we skipped the information regarding the authors) to create their textual representation.

The task in those experiments was to learn how to tag the documents from Infona with the categories from the OECD FOS classification. This classification system consists of 42 main categories grouped into six different upper-level categories, namely *Natural sciences*, *Engineering and technology*, *Medical and Health sciences*, *Agricultural sciences*, *Social sciences* and *Humanities* [17]. In order to construct

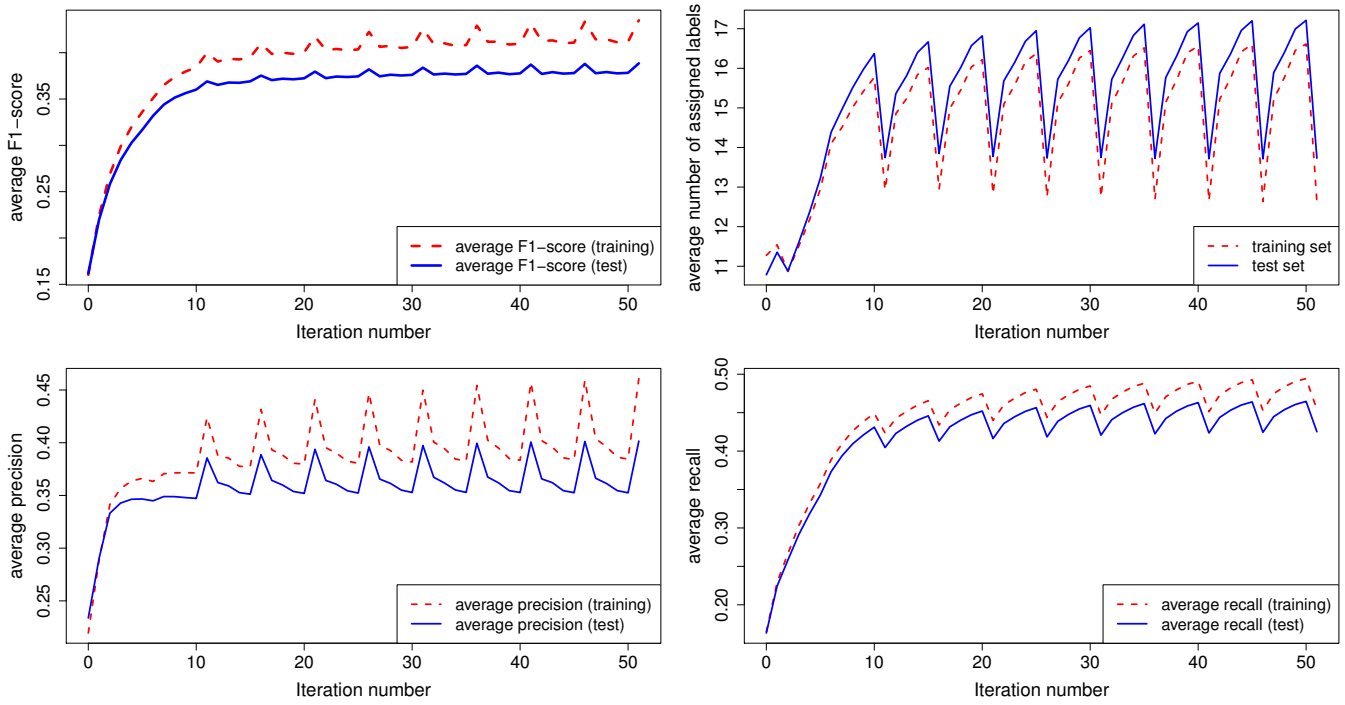

Fig. 3. Results of the adaptive learning algorithm on the PubMed Central corpus (50 iterations). The documents were labeled with the concepts from the MeSH ontology. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

the textual descriptions for all the 42 categories we manually selected a number of related English Wikipedia articles and we concatenated their content.

The size of the corpus for those experiments was limited by the availability of the expert knowledge. Infona did not provide us any information about the OECD FOS categories of the documents. In order to create a reference set of labeled documents we had to ask volunteers to manually tag the data. In this way we obtained a total of 1000 labeled meta-data entries which we could use in the experiments. A single document on average was assigned to 1.7 categories. Approximately 30% of the documents were labeled by more than one person. In this way we could check how difficult this task is and get a good estimation of a reference quality assessment. It turned our that the average cross-expert F_1-$score$ merely exceeds 0.51, while the average $Precision$ and $Recall$ values are about 0.55. It means that on average, two different experts agree only on about a half of the assigned categories, thus we should not expect a better result from an automatic tagging method.

In the experiments we used 800 documents as a training set and the remaining 200 served as a test set. Due to the small size of the test sets, we repeated the testing procedure 20 times on different divisions of the data in order to get reliable estimations of the tagging quality. Similarly as in the case of the ACM Digital Library corpus, we set the initial values of the activation thresholds to a low value, i.e. they were all equal 0.25. The average results of those tests for the consecutive iterations of the ASA algorithm are presented in Figure 5. In those plots, the values of the cross-expert quality estimations are marked by the thick black lines.

The experimental results once again clearly demonstrate usefulness of our learning algorithm. The average F_1-$score$ value obtained using the adapted inverted index was greater by over 40% than the score of the standard ESA algorithm. For ASA it was approximately 0.47. It is worth noting that this improvement was possible, even though the number of available training documents was very limited. The best F_1-$score$ on the test set was usually achieved around thirtieth iteration of the algorithm and after that point we noted a slight decrease in the results. Since the scores obtained on the training set systematically grew and often exceeded 0.8, this can be most likely explained by the over-fitting problem. Nevertheless, the score achieved using ASA was very close to the cross-expert F_1-$score$ which confirms the effectiveness of the proposed adaptive learning algorithm.

V. Conclusions

In the paper we discussed an adaptive learning framework, called Adaptive Semantic Analysis, which can be utilized for improving the terms-to-concepts associations from the inverted index of the ESA algorithm. We described in details the learning procedure and we showed its effectiveness in dealing with real-life problems. In particular, we presented results of experiments on three document corpora, in which the ASA algorithm was used to facilitate the automatic tagging of documents with concepts from different knowledge bases.

We hope that in a future our automatic tagging module can become a part of a larger scientific article repository platform. We are currently trying to integrate our SONCA platform with the Infona repository. This may enable an

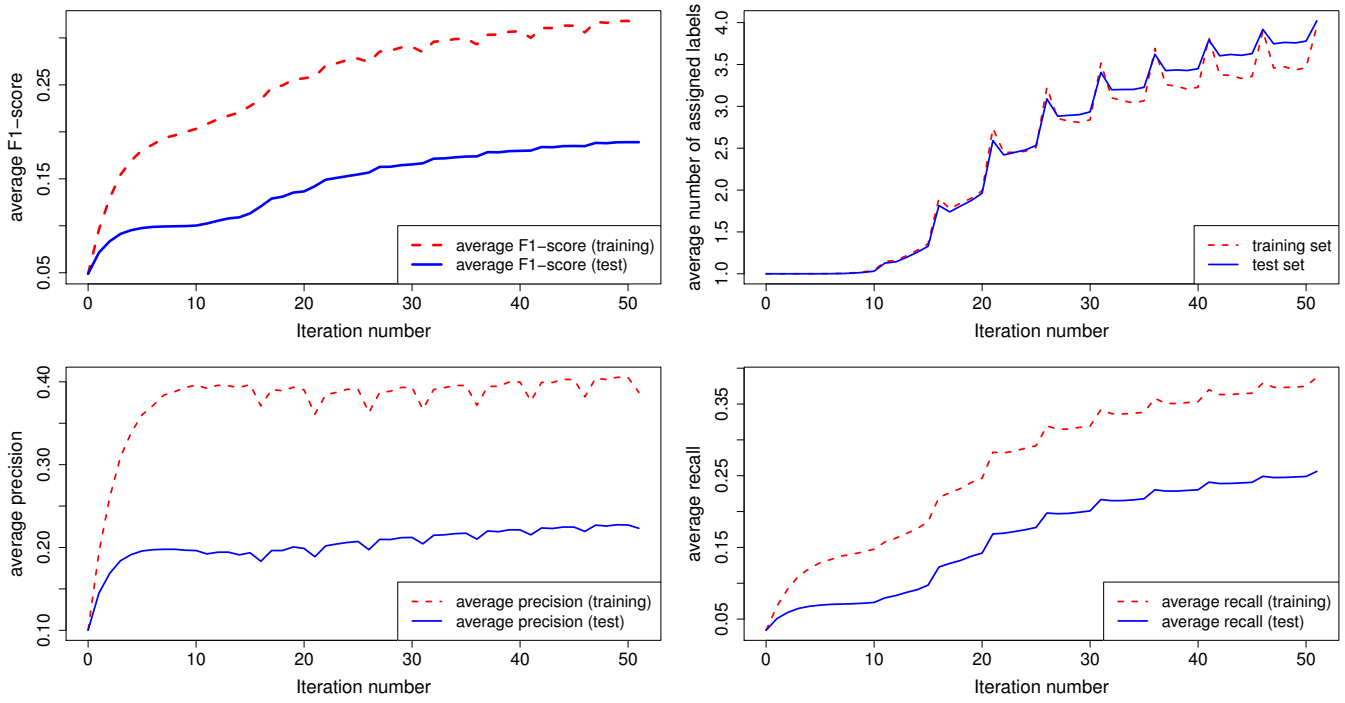

Fig. 4. Results of the adaptive learning algorithm on the ACM Digital Library corpus (50 iterations). The documents were labeled with the ACM Computing Classification System categories. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

efficient and automatic semantic indexing of Infona's resources which would allow to better fulfill the information needs of Infona's users. Apart from the direct use as an indexing module of the search engine, the tags returned by our system could be utilized for, e.g., improving the clustering of search results or assigning comprehensible names to document clusters.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[2] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz, "Semantic web search based on ontological conjunctive queries," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2011.

[3] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," *Int. Journal on Semantic Web and Information Systems (IJSWIS). Special Issue of Multimedia Semantics*, vol. 3, no. 3, pp. 55–73, 2006.

[4] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron, "Semantic analytics of PubMed content," in *Information Quality in e-Health - 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2011, Graz, Austria, November 25-26, 2011. Proceedings*, ser. LNCS, A. Holzinger and K.-M. Simonic, Eds., vol. 7058. Springer, 2011, pp. 63–74.

[5] A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the web," *ACM Trans. Internet Technol.*, vol. 9, pp. 1–24, 2009. [Online]. Available: http://doi.acm.org/10.1145/1552291.1552293

[6] R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka, Eds., *Intelligent Tools for Building a Scientific Information Platform - Advanced Architectures and Solutions*, ser. Studies in Computational Intelligence. Springer, 2013, vol. 467.

[7] L. A. Nguyen and H. S. Nguyen, "On designing the sonca system," in *Intelligent Tools for Building a Scientific Information Platform*, R. Bembenik, L. Skonieczny, H. Rybiński, and M. Niezgódka, Eds. Springer-Verlag New York, 2012, pp. 9–36.

[8] A. Janusz, W. Świeboda, A. Krasuski, and H. S. Nguyen, "Interactive document indexing method based on Explicit Semantic Analysis," in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 156–165.

[9] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. of The 20th Int. Joint Conf. on Artificial Intelligence*, Hyderabad, India, 2007, pp. 1606–1611. [Online]. Available: http://www.cs.technion.ac.il/~shaulm/papers/pdf/Gabrilovich-Markovitch-ijcai2007.pdf

[10] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, pp. 8:1–8:34, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1961209.1961211

[11] United States National Library of Medicine, "Introduction to MeSH - 2011," Online: http://www.nlm.nih.gov/mesh/introduction.html, 2011. [Online]. Available: http://www.nlm.nih.gov/mesh/introduction.html

[12] Association for Computing Machinery, "The 2012 acm computing classification system," Online: http://www.acm.org/about/class/2012, 2012. [Online]. Available: http://www.acm.org/about/class/2012

[13] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[15] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski, "JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers," in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 417–426.

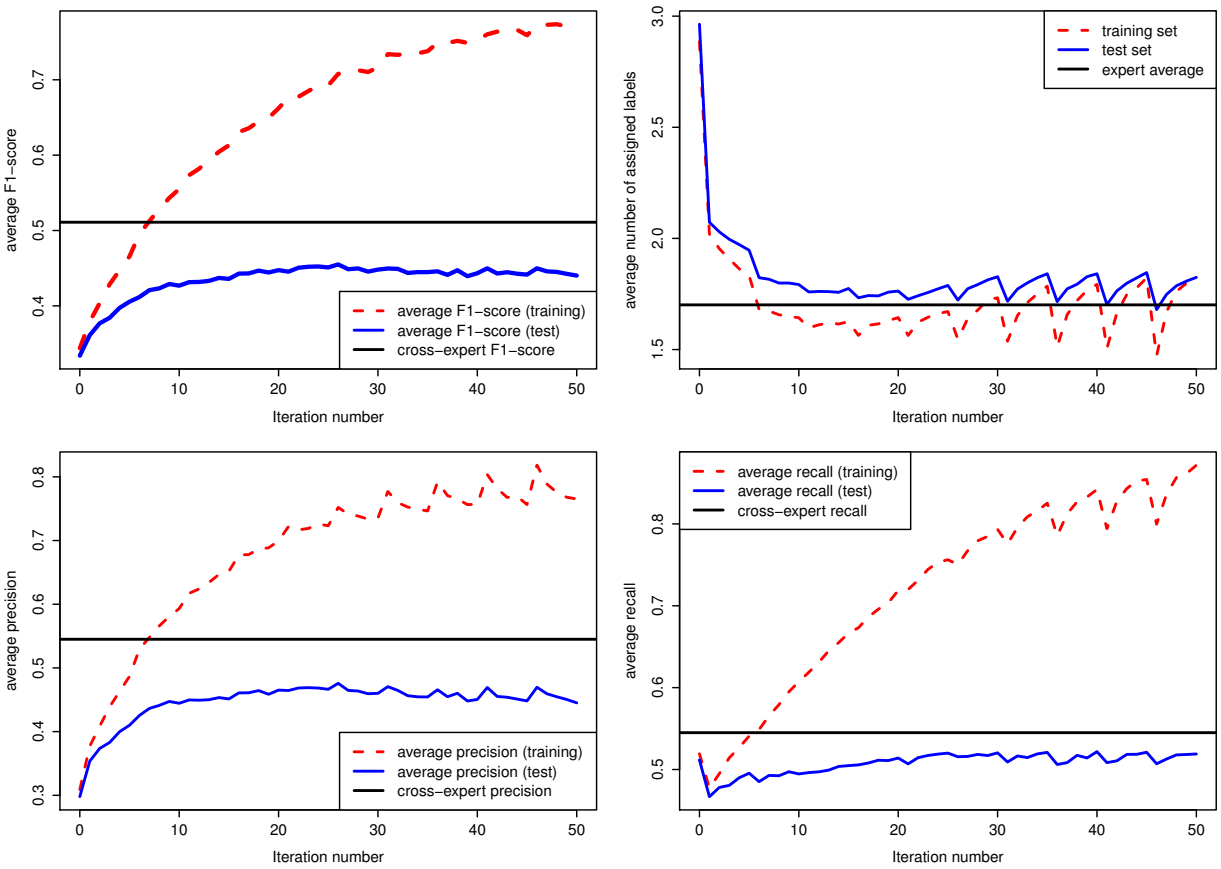

Fig. 5. Results of the adaptive learning algorithm on the corpus from the Infona repository (50 iterations). The documents were labeled with the OECD Fields of Science and Technology Classification categories. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. Moreover, the thick black line in the plots denotes the cross-expert statistic values which can be regarded as an additional reference. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

[16] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 381–382, 2001. [Online]. Available: http://www.pnas.org/content/98/2/381.abstract

[17] "Revised Field of Science and Technology (FoS) Classification in the Frascati Manual," Committee for Scientific and Technological Policy, Directorate for Science, Technology and Industry, OECD, Feb. 2007.