Automatic Web Service Tagging Using Machine Learning and WordNet Synsets

Zeina Azmeh, Jean-Rémy Falleri, Marianne Huchard, Chouki Tibermacine

Automatic Web Service Tagging Using Machine Learning and WordNet Synsets*

Zeina Azmeh, Jean-Rémy Falleri, Marianne Huchard and Chouki Tibermacine

LIRMM, CNRS and Montpellier II University
161, rue Ada 34392 Montpellier Cedex 5 France
{falleri,azmeh,huchard,tibermacin}@lirmm.fr

Abstract. *The importancy of Web services comes from the fact that they are an important means to realize SOA applications. Their increasing popularity caused the emergence of a fairly huge number of services. Therefore, finding a particular service among this large service space can be a hard task. User tags have proven to be a useful technique to smooth browsing experience in large document collections. Some service search engines proposes the facility of service tagging. It is usually done manually by the providers and the users of the services, which can be a fairly tedious and error prone task. In this paper we propose an approach for tagging Web services automatically. It adapts techniques from text mining and machine learning to extract tags from WSDL descriptions. Then it enriches these tags by extracting relevant synonyms using WordNet. We validated our approach on a corpus of 146 services extracted from Seekda.*

Keywords: Tags, web services, text mining, machine learning.

1 Introduction

Service-oriented architectures (SOA) are achieved by connecting loosely coupled units of functionality. The most common implementation of SOA uses units of functionality invokable though Internet, called Web Services. Using SOA, a developer can quickly build a complex software by using already available Web services. One of the main tasks is therefore to find the relevant Web services to use in the software. With the increasing interest toward SOA, the number of existing Web services is dramatically growing. Finding a particular service among this huge amount of services is becoming a time-consuming task.

Web services are usually described with a standard XML-based language called WSDL. The WSDL format has been designed to be processed automatically by programs, but it includes a documentation part that can be filled with a text indicating to the user what the service do. Unfortunately, this documentation part is often not filled by the creators of the services. In this case, the potential users of the service spend time to understand its functionality and to

* France Télécom R&D has partially supported this work (contract CPRE 5326).

decide whether or not they will use it. Moreover, it is common that a user has finally selected a service but finds out that in fact this service is irrelevant. When this case occurs, the user might want to easily get a list of services offering a similar functionality. *Tagging* is a mechanism that has been introduced in search engines and digital libraries to fulfill exactly this objective.

Tagging is the process of describing a resource by assigning some relevant keywords (tags) to it. The tagging process is usually done manually by the users of the resource to be tagged. Tags are useful when browsing large collections of documents. Indeed, unlike with traditional hierarchical categories, documents can be assigned an unlimited number of tags. It allows cross-browsing between the documents. Seekda[1], one of the main service search engines, already allows its users to tag its indexed services. Tags are also useful to have a quick understanding of a particular service. Moreover, since tags are words particularly important for the services, they are a good basis for other important tasks, like service classification or clustering.

In this paper, we present an approach that automatically extract a set of relevant tags from a WSDL service description, documented or not. We use a corpus of user-tagged services to learn how to extract relevant tags from untagged service descriptions. Our approach relies on text mining techniques in order to extract candidate tags out of a description, and machine learning techniques to select relevant tags among these candidates. The extracted set of tags is then enriched with semantically related tags using the WordNet ontology [11]. We have validated this approach on a corpus of 146 user-tagged Web services extracted from Seekda. Results show that this approach is significantly more efficient than the traditional (but fairly efficient) *tfidf* weight (explained in Section 2.1).

The remaining of the paper is organized as follows. Section 2 introduces the context of our work. Then, Section 3 details our tag extraction process. Section 4 presents a validation of this process and discusses the obtained results. Before concluding and presenting the future work, we describe the related work in Section 5.

2 Context of the Work

Our work focuses on extracting tags from service descriptions. In the literature, we found a similar problem: *keyphrase extraction*. Keyphrase extraction aims at extracting important and relevant short phrases from a plain-text document. It is mostly used on journal articles or on scientific papers in order to smooth browsing and indexation of those documents in digital libraries. Before starting our work, we analyzed one assessed approach that performs keyphrase extraction: Kea [12] (Section 2.1). After this analysis, we concluded that a straightforward application of this approach is not possible on service descriptions instead of plain-text documents (Section 2.2).

[1] http://www.seekda.com

2.1 Description of Kea

Kea [12] is a keyphrase extractor for plain-text documents. It uses a Bayesian classification approach. Kea has been validated on several corpora [16,17] and has proven to be an efficient approach. It takes a plain-text document as input. From this text, it extracts a list of candidate keyphrases. These candidates are the $\bigcup_{i=1}^{k} k$-grams of the text. For instance, let us consider the following sample document: "*I am a sample document*". The candidate keyphrases extracted if $k = 2$ are: *(I,am,a,sample,document,I am,am a,a sample,sample document)*. To choose the most adapted value of k for the particular task of extracting tags from WSDL files, we made some measurments and found that 86% of the tags are of length 1. It clearly shows that one word tags are assigned in the vast majority of the cases. Therefore we will fix $k = 1$ in our approach (meaning that we are going to find one word length tags). Nevertheless, our approach, like Kea, is easily generalizable to extract tags of length k.

Kea then computes two features on every candidate keyphrase. First, *distance* is computed, which is the number of words that precede the first observation of the candidate divided by the total number of words of the document. For instance, for the sample document, $distance(am\ a) = \frac{1}{5}$. Second, *tfidf*, a standard weight in the information retrieval field, is computed. It measures how much a given candidate keyphrase of a document is specific to this document. More formally, for a candidate c in a document d, $tfidf(c,d) = tf(c,d) \times idf(c)$. The metric $tf(c,d)$ (*term frequency*) corresponds to the frequency of the term c in d. It is computed with the following formula: $tf(c,d) = \frac{occurences\ of\ c\ in\ d}{size\ of\ d}$. The metric $idf(c)$ (*inverse document frequency*) measure the general importance of the term in a corpus $\mathcal{D}$. $idf(c) = log(\frac{|\mathcal{D}|}{|\{d:\ c\ \in d\}|})$.

Kea uses a naive Bayes classifier to classify the different candidate keyphrases using the two previously described features. The authors showed that this type of classifier is optimal [7] for this kind of classification problem. The two classes in which the candidate keyphrases are classified are: *keyphrase* and *not keyphrase*. Several evaluations on real world data report that Kea achieve good results [16,17]. In the next section, we will describe how WSDL files are structured and highlight why the Kea approach is not directly applicable on this kind of data.

2.2 WSDL service descriptions

The documents from which we intend to extract tags are service descriptions in the WSDL format. This format is XML-based and aims at describing the different elements involved in a web service. Those elements are: *services*, *ports*, *port types*, *bindings*, *types* and *messages*. Their descriptions always come with a name, called *identifier* (example: *MyWeatherService*, *ComputeExchangeRatePort*). They can optionally come with a plain-text documentation. Figure 3 (left) shows the general outline of a WSDL file.

One simple idea to extract tags from services would be to use Kea on their plain-text documentations. Unfortunately, an analysis of our service corpus (see

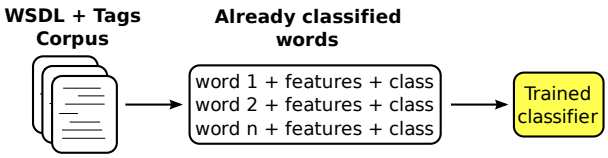

Fig. 1. The training phase

Section 3.1 for more information about this corpus) shows that about 38% of the services are not documented at all. Using only plain-text documentation of the WSDL to tags service would therefore leaves at least 38% of the services untagged, which is not acceptable. Another important source of information to discover tags are the identifiers contained in the WSDL. For instance *weather* would surely be an interesting tag for a service named *WeatherService*. Unfortunately, identifiers are not easy to work with. Firstly because identifiers are usually a concatenation of different words (remember $MyWeatherService$). Secondly because they can be associated to different kinds of elements (services, ports, types, ...) that have not the same importance in a service description. For all of these reasons, extracting candidate tags from WSDL files is not straightforward. Several pre-processing and text-mining techniques are required. Moreover, the previous described feature (*tfidf* and *distance*) are not easy to adapt on words coming from a service descriptions. First because WSDL deals with two categories of words (the one coming from the documentation and the one coming from the identifiers) that are not necessary related. Second because the *distance* feature is meaningless on the identifiers, which are defined in an arbitrary order.

3 Tag Extraction Process

Similarly to Kea, we model the tag extraction problem as the following classification problem: classifying a word into one of the two *tag* and *no tag* classes. Our overall process is divided into two phases: the *training* phase and the *tag extraction* phase.

Figure 1 summarizes the behavior of the training phase. In this phase we dispose of a corpus of WSDL files and associated tags. Since we did not find such a publicly available corpus, we created one by using data from Seekda. The creation of this *training corpus* is described in Section 3.1. From this training corpus, we first extract a list of candidate words by using text-mining techniques. The extraction of these candidates is described in Sections 3.2 and 3.3. Then several *features* are computed on every candidate. A *feature* is a common term in the machine learning field. It can be seen as an attribute that can be computed on the candidates (for instance the frequency of the words in their WSDL file). Finally, since manual tags are assigned to those WSDL files, we use them to classify the candidate words coming from our WSDL files. Using this set of candidate words, computed features and assigned classes, we train a classifier.

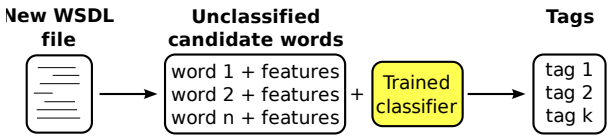

Fig. 2. The tag extraction phase

This trained classifier will then be used to classify words coming from subsequent WSDL files during the *tag extraction* phase.

Figure 2 describes the *tag extraction phase*. First, like in the *training phase*, a list of candidate words is extracted from an untagged WSDL file. The same features as in the training phase are then computed on those words. The only difference with the training phase is that we do not know in advance which of those candidates are true tags. Therefore we use the previously trained classifier to automatically perform this classification. Finally the tags extracted from the WSDL file are the words that have been classified in the *tag* class. It is noteworthy to remark that the *training* phase is only performed once, while the *tag extraction* phase can be applied an unlimited number of times.

3.1 Creation of the training corpus

As explained above, our approach requires a training corpus, denoted by $\mathcal{T}$. Since we want to extract tags from WSDL files, $\mathcal{T}$ has to be a set of couples $(wsdl, tags)$, with $wsdl$ a WSDL file, and $tags$ a set of corresponding manually assigned tags. We were not aware of such a publicly available corpus. Therefore we decided to create one using data from Seekda. Indeed, Seekda allows its users to manually assign tags to its indexed services. We created a program that crawls on the Seedka services and extracts the WSDL files together with the user tags. To ensure that the services of our corpus were significantly tagged, we only retain the WSDL files that have at least five tags. Using this program, we extracted 150 WSDL files. Then, we removed from $\mathcal{T}$ the WSDL files that triggered parsing errors. Finally, we dispose of a training corpus containing 146 WSDL files together with their associated tags.

To clean the tags of the training corpus, we performed the three following operations:

- We removed the non alpha numeric characters from the tags (we found several tags like *_onsale* or *:finance*),
- We removed a meaningless and highly frequent tag (the *_unkown* tag),
- We divided the tags with length $n > 1$ into n tags of length 1, in order to have only tags of length 1 (the reason has been explained in section 2.1). The length of a tag is defined as the number of words composing this tag.

Finally, we dispose of a corpus of 146 WSDL files and 1393 tags (average of 9.54 tags per WSDL). An analysis of $\mathcal{T}$ shows that about 35% of the user tags are already contained in the WSDL files. Now that we have this training corpus, we will shortly describe the approach upon which our work is built.

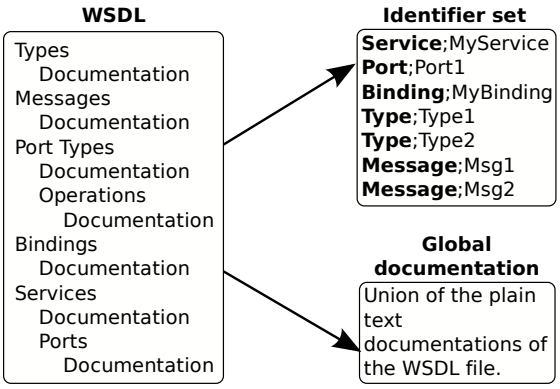

Fig. 3. WSDL pre-processing

3.2 Pre-processing of the WSDL files

As we have seen before, a WSDL file contains several element definitions optionally containing a plain-text documentation. The left side of figure 3 shows such a data structure. In order to simplify the WSDL XML representation in a format more suitable to apply text mining techniques, we decided to extract two documents from a WSDL description:

- A set of couples $(type, ident)$ representing the different elements defined in the WSDL. We have $type \in$ *(Service,Port,PortType,Message,Type,Binding)* the type of the element and *ident* the identifier of the element. We call this set of couples the *identifier set*.
- A plain text containing the union of the plain-text documentations found in the WSDL file, called the *global documentation*.

This pre-processing operation is summarized in the figure 3.

3.3 Selection of the candidate tags

As seen in the previous section, we dispose now of two different sources of information for a given WSDL: an *identifier set* and a *global documentation*. Unfortunately, those data are not yet usable to compute meaningful metrics. Firstly because the identifiers are names of the form *MyWeatherService*, and therefore are very unlikely to be tags. Secondly because this data contains a lot of obvious useless tags (like the *you* pronoun). Therefore, we will now apply several text-mining techniques on the identifier set and the global documentation.

Figure 4 shows how we process the *identifier set*. Here is the complete description of all the performed steps:

1. **Identifier type filtering**: during this step, the couples *(type,ident)* where $type \in$ *(PortType,Message,Binding)* are discarded. We applied this filtering because very often, the identifiers of the elements in those categories are duplicated from the identifiers in the others categories.

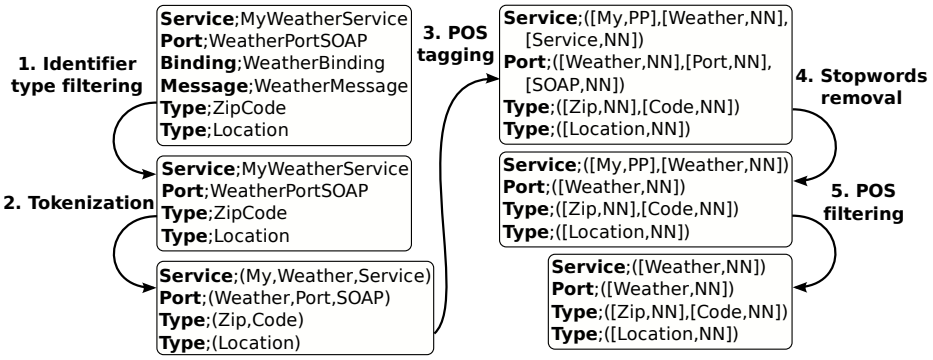

Fig. 4. Processing of the identifiers

2. **Tokenization**: during this step, each couple $(type, ident)$ is replaced by a couple $(type, tokens)$. $tokens$ is the set of words appearing in $ident$. For instance, *(Service,MyWeatherService)* would be replaced by *(Service,[My,Weather,Service])*. To split $ident$ into several tokens, we created a tokenizer that uses common clues in software engineering to split the words. Those clues are for instance a case change, or the presence of a non alpha-numeric character.
3. **POS tagging**: during this step each couple $(type, tokens)$ previously computed is replaced by a couple $(type, ptokens)$. $ptokens$ is a set of couples $(token_i, pos_i)$ derived from $tokens$ where $token_i$ is a token from $tokens$ and pos_i the part-of-speech corresponding to this token. We used the tool *tree tagger* [25] to compute those part-of-speeches. Example: *(Service,[My,Weather,Service])* is replaced by *(Service,[(My,PP),(Weather,NN),(Service,NN)])*. *NN* means noun and *PP* means pronoun.
4. **Stopwords removal**: during this step, we process each couple $(type, ptokens)$ and remove from $ptokens$ the elements $(token_i, pos_i)$ where $token_i$ is a *stopword* for $type$. A stopword is a word too frequent to be meaningful. We manually established a stopword list for each identifier type. Example: *(Service,[(My,PP),(Weather,NN),(Service,NN)])* is replaced by *(Service,[(My,PP)(Weather,NN)])* because *Service* is a stopword for service identifiers.
5. **POS filtering**: during this step, we process each couple $(type, ptokens)$ and remove from $ptokens$ the elements $(token_i, pos_i)$ where $pos_i \notin$ *(Noun,Adjective,Verb,Symbol)*. Example: *(Service,[(My,PP),(Weather,NN))* is replaced by *(Service,[(Weather,NN)])* because pronouns are filtered.

Figure 5 shows how we process the *global documentation*. Here is the complete description of all the performed steps:

1. **HTML tags removal**: the HTML tags (words begining by $<$ and ending by $>$) are removed from the global documentation.
2. **POS tagging**: similar to the POS tagging step applied to the identifier set.
3. **POS filtering**: similar to the POS filtering step applied to the identifier set.

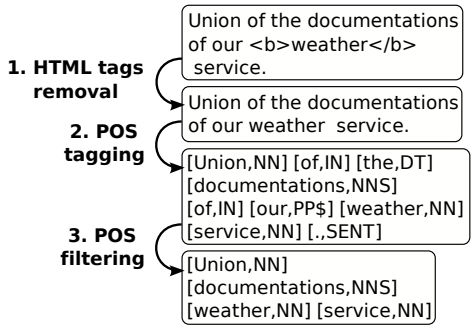

Fig. 5. Processing of the global documentation

The union of the remaining words in the identifier set and in the global documentation are our candidate tags. When defining those processing operations, we took great care that no correct candidate tags (i.e. a candidate tag that is a real tag) of the training corpus have been discarded. The next section describes how we adapted the Kea features to these candidate tags.

3.4 Computation of the features

After having applied our text mining techniques on the identifier set and the global documentation, we dispose now of different well separated words. Therefore we can now compute the $tfidf$ feature. But words appearing in documentation or in the identifier names are not the same. We decided (mostly because it turns out to perform better) to separate the $tfidf$ value into a $tfidf_{ident}$ and a $tfidf_{doc}$ which are respectively the $tfidf$ value of a word over the identifier set and over the global documentation. Like in Kea, we used the method in [10] to discretize those two real-valued features.

The *distance* feature still has no meaning over the identifier set, because the elements of a WSDL description are given in an arbitrary order. Therefore we decided to adapt it by defining five different features: *in_service*, *in_port*, *in_type*, *in_operation* and *in_documentation*. Those features take their values in the $(true, false)$ set. A *true* value indicates that the word has been seen in an element identifier of the corresponding type. For instance *in_service(weather) = true* means that the word *weather* has been seen in a service identifier. *in_documentation(weather) = true* means that the word *weather* has been seen in the global documentation.

In addition of these features, we compute another feature called *pos*. We added this feature, not used in Kea, because it significantly improves the results. *pos* is simply the part-of-speech that has been assigned to the word during the POS tagging step. If several parts-of-speech have been assigned to the same word, we choose the one that has been assigned in the majority of the cases. The different values of *pos* are: *NN (noun), NNS (plural noun), NP (proper noun),*

Table 1. Extract of the ARFF file

Word	$TFIDF_{id}$	$TFIDF_{doc}$	$IN\_SERVICE$	...	IN_DOC	POS	IS_TAG
Weather	$[0, 0.01]$	$]0.01, 0.04]$	×			NN	×
Location	$]0.03, 0.1]$	$]0.04, 0.15]$			×	JJ	
Code	$]0.03, 0.1]$	$]0.01, 0.04]$			×	VV	

NPS (plural proper noun), JJ (adjective), JJS (plural adjective), VV (verb), VVG (gerundive verb), VVD (preterit verb), SYM (symbol).

3.5 Training and using the classifier

We applied the previously described technique to all the WSDL files of $\mathcal{T}$. In addition to the previously described features, we compute the is_tag feature over the candidates. This feature takes its values in the $(true, false)$ set. $is\_tag(word) = true$ means that $word$ has been assigned as a tag by Seekda users for its service description. We have serialized all those results in an *ARFF* file compatible with the Weka tool [29]. Weka is a machine learning tool that defines a standard format for describing a training corpus and furnish the implementation of many classifiers. One can use Weka in order to train a classifier or compare the performances of different classifiers regarding a given classification problem. Table 1 shows an extract of the ARFF file we produce. In this table, words are displayed for the sake of clarity, but in reality, they are not present in the ARFF file. The ARFF file only contains features.

With this ARFF file, we used Weka to train a naive Bayes classifier, shown as optimal for our kind of classification task [7]. This trained classifier can now be used in the tag extraction phase. As previously said, the beginning of this phase is the same as the one of the training phase. It means that the WSDL file goes through the previously described operations (pre-processing, candidates selection and features computation). Only this time, the value of the is_tag feature is not available. This value will be automatically computed by the previously trained classifier.

3.6 WordNet for semantically related tags

In our approach, the classifier that we built determines whether a word in a WSDL file is a tag or not. Thus, it extracts the tags appearing inside the WSDL files only. This way, we miss some other interesting tags like associated words or synonyms. In order to solve this issue, we used the WordNet lexical database [11]. In WordNet a word may be associated with many synsets (synonym sets), each corresponding to a different sense of a word.

Our corpus consists of 146 WSDL files, each of which is assigned two sets of tags: user tags and our automatically extracted tags. Our objective is to enrich each set of tags with semantically similar words extracted from WordNet. Thus, for each tag we identify the possible senses and the synonyms set related to each sense. We add the extracted synonyms to the corresponding set of tags, and we

perform some experiments to evaluate the obtained tags, as we show in the next section.

4 Validation of the Proposed Work

This section provides a validation of our technique on real world data from Seekda. We conduct experiments in which we assess the precision and recall of our trained classifier.

Methodology: We carried out our experiments on three stages. In the first one, the trained classifier is applied on the training corpus $\mathcal{T}$ and its output is compared with the tags given by Seekda users (obtained as described in Section 3.1).

After having conducted the first experiment, a manual assessment of the tags produced by our approach revealed that many tags not assigned by the user seemed highly relevant. This phenomenon has also been observed in several human evaluations of Kea [16,17], that inspired our approach. It occurs because tags assigned by the users are not the *absolute truth*. Indeed, it is very likely that users have forgotten many relevant tags, even if they were in the service description. To show that the real efficiency of our approach is better than the one computed in the first experiment, we perform a second experiment. In this experiment, we manually augmented the user tags of our corpus with additional tags we found relevant and accurate by analyzing the WSDL descriptions of the services. In the final experiment, we enriched the user tags as well as our automatically extracted tags with semantically related tags using WordNet.

Metrics: In the evaluation, we used precision and recall. First, for each web service $s \in \mathcal{T}$, where $\mathcal{T}$ is our training corpus, we consider: A the set of tags produced by the trained classifier, M the set of the tags given by Seekda users and W the set of words appearing in the WSDL. Let $I = A \cap M$ be the set of tags assigned by our classifier and Seekda users. Let $E = M \cap W$ be the set of tags assigned by Seekda users present in the WSDL file. Then we define $precision(s) = \frac{|I|}{|A|}$ and $recall(s) = \frac{|I|}{|E|}$, which are aggregated in $precision(\mathcal{T}) = \frac{\sum_{s \in \mathcal{T}} precision(s)}{|\mathcal{T}|}$ and $recall(\mathcal{T}) = \frac{\sum_{s \in \mathcal{T}} recall(s)}{|\mathcal{T}|}$. The recall is therefore computed over the tags assigned by Seekda users that are present in the descriptions of concerned services. We did not compute the recall for the WordNet extracted tags, because these tags may not be present in the WSDL descriptions.

Evaluation: Figure 6 (left) gives results for the first experiment where the output of the classifier is compared with the tags of Seekda users, while in Figure 6 (right), enriched tags of Seekda users are used in the comparison (curated corpus). In this figure, our approach is called *ate* (*Automatic Tag Extraction*). To clearly show the concrete benefits of our approach, we decided to include in these experiments a straightforward (but fairly efficient) technique. This technique,

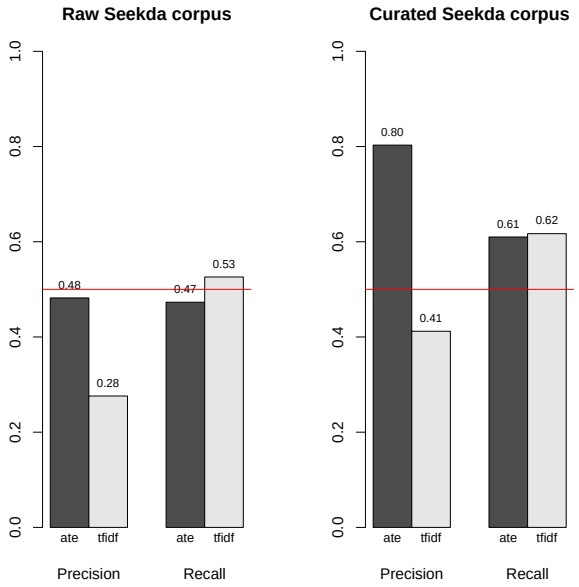

Fig. 6. Results on the original and manually curated Seekda corpus

called *tfidf* in Figure 6, consists in selecting, after the application of our text-mining techniques, the five candidate tags with the highest *tfidf* weight.

In Figure 6 (left), the precision of *ate* is 0.48. It is a significant improvement compared to the *tfidf* method that achieves only a precision of 0.28. Moreover, there is no significant difference between the recall achieved by the two methods. To show that the precision and recall achieved by *ate* are not biased by the fact that we used the training corpus as a testing corpus, we performed a 10 folds cross-validation. In a 10 folds cross-validation, our training corpus is divided in 10 parts. One is used to train a classifier, and the 9 other parts are used to test this classifier. This operation is done for every part, and then, the average recall and precision are computed. The results achieved by our approach using cross-validation ($precision = 0.44$ and $recall = 0.42$) are very similar to those obtained in the first experiment.

In Figure 6 (right), we see that the precision achieved by *ate* in the second experiment is much better. It reaches 0.8, while the precision achieved by the *tfidf* method increases to 0.41. The recall achieved by the two methods remains similar. The precision achieved by our method in this experiment is good. Only 20% of the tags discovered by *ate* are not correct. Moreover, the efficiency of *ate* is significantly higher than *tfidf*.

Evaluation after using WordNet: We enriched the tags sets with semantically similar words extracted using the WordNet, as described above. We recal-

culated the precision value, considering these new sets of enriched user tags and automatically extracted tags. The precision value has increased by 9%, reaching the value of 89% of correctness. Thus, using the WordNet has improved the precision value and enriched the services with tags that are not necessarily present in the WSDL descriptions.

Threats to validity: Our experiments use real world services, obtained from the Seekda service search engine. Our training corpus contains services extracted randomly with the constraint that they contain at least 5 user tags. We assumed that Seekda users assign correct tags. Indeed, our method admits some noise but would not work if the majority of the user tags were poorly assigned. In the second experiment, we manually added tags we found relevant by examining the complete description and documentation of the concerned services. Unfortunately, since we are not "real" users of those services, some of the tags we added might not be relevant.

5 Related Work

In this section, we will present the related work according to two fields of research: keyphrase extraction and web service discovery.

5.1 Keyphrase extraction and assignment

According to [27], there are two general approaches that are able to supply keyphrases for a document: *keyphrase extraction* and *keyphrase assignment*. Both approaches are using supervised machine learning approaches, with training examples being documents with manually supplied keyphrases.

Keyphrase assignment: In the *keyphrase assignment* approach, a list of predefined keyphrases is treated as a list of classes in which the different documents are classified. Text categorization techniques are used to learn models for assigning a class (*keyphrase*) to a document. Two main approaches of this category are [9,19].

Keyphrase extraction: In the *keyphrase extraction* approach, a list of candidate keyphrases are extracted from a document and classified into the classes *keyphrase* and *not keyphrase*. There are two main approaches that fall in this category: one using a genetic algorithm [26] and one using a naive Bayes classifier (Kea [12]).

5.2 Web service discovery

Web service discovery is a wide research area with many underlying issues and challenges. A quick overview of some of the works can be acquired from [4,18][2]. Here, we describe a selection of works, classified using their adapted techniques.

[2] The second one is edited by the responsible of Seekda's technical infrastructure

Using machine learning techniques: Many approaches adapt techniques from machine learning field, in order to discover and group similar services. In [5,15], service classifiers are defined depending on sets of previously categorized services. Then the resulting classifiers are used to deduce the relevant categories for new given services. In case there were no predefined categories, unsupervised clustering is used. In [21], CPLSA approach is defined that reduces a services set then cluster it into semantically related groups.

Using service matching techniques: In [20], a web service broker is designed relying on approximate signature matching using xml schema matching. It can recommend services to programmers in order to compose them. In [13], a service request and a service are represented as two finite state machines then they are compared using various heuristics to find structural similarities between them. In [8], the Woogle web service search engine is presented, which takes the needed operation as input and searches for all the services that include an operation similar to the requested one. In [3], tags coming from folksonomies are used to discover and compose services.

Using vector space model techniques: The vector space model is used for service retrieval in several existing works as in [23,28,6]. Terms are extracted from every WSDL file and the vectors are built for each service. A query vector is also built, and similarity is calculated between the service vectors and the query vector. This model is sometimes enhanced by using WordNet, structure matching algorithms to ameliorate the similarity scores as in [28], or by partitioning the space into subspaces to reduce the searching space as in [6].

Using formal concept analysis techniques: A collection of works [1,22,2], adapt the formal concept analysis method to retrieve web services more efficiently. Contexts obtained from service descriptions are used to classify the services as a concept lattice. This lattice helps in understanding the different relationships between the services, and in discovering service substitutes.

6 Conclusion and Future Work

With the emergence of SOA, it becomes important for developers using this paradigm to retrieve Web services matching their requirements in an efficient way. By using Web service search engines, these developers can either search by keywords or navigate by tags. In the second case, it is necessary that the tags characterize accurately this service. Our work contributes in this direction and introduces a novel approach that extracts tags from Web service descriptions and enrich them using the WordNet ontology. This approach combines and adapts text mining as well as machine learning techniques. It has been experimented on a corpus of user-tagged real world Web services. The obtained results demonstrated the efficiency of our automatic tag extraction process, and the enrichment of semantically similar tags. The use of WordNet has improved the

precision of the returned tags and enriched the services with tags that are not necessarily present in the WSDL descriptions. The proposed work is useful for many purposes. First, the automatically extracted tags can assist the users who are tagging a given service, or to "bootstrap" tags on untagged services. They are also useful to have a quick understanding of a service without reading the whole description. They can also be used to help in building domain ontologies like in [24] [14], also in tasks such as service clustering (for instance by measuring the similarity of the tags of two given services), or classification (for instance by defining association rules between tags and categories). One of our perspectives is to work on extracting composed tags, which consist of more than one word. A one-word tag is sometimes insufficient to describe some concepts (for example *exchange rate* or *Web 2.0*).

References

1. Aversano, L., Bruno, M., Canfora, G., Penta, M.D., Distante, D.: Using concept lattices to support service selection. International Journal of Web Services Research 3(4), 32–51 (2006)
2. Azmeh, Z., Huchard, M., Tibermacine, C., Urtado, C., Vauttier, S.: Wspab: A tool for automatic classification & selection of web services using formal concept analysis. In: Proceedings of the 6th IEEE European Conference on Web Services (ECOWS 2008). pp. 31–40. IEEE Computer Society, Dublin, Ireland (2008)
3. Bouillet, E., Feblowitz, M., Feng, H., Liu, Z., Ranganathan, A., Riabov, A.: A folksonomy-based model of web services for discovery and automatic composition. In: IEEE International Conference on Services Computing (SCC). pp. 389–396. IEEE Computer Society (2008)
4. Brockmans, S., Erdmann, M., Schoch, W.: Service-finder deliverable d4.1. research report about current state of the art of matchmaking algorithms. Tech. rep., Ontoprise, Germany (October 2008)
5. Crasso, M., Zunino, A., Campo, M.: Awsc: An approach to web service classification based on machine learning techniques. Inteligencia Artificial, Revista Iberoamericana de Interligencia Artificial 12, No 37, 25–36 (2008)
6. Crasso, M., Zunino, A., Campo, M.: Query by example for web services. In: SAC '08: Proceedings of the 2008 ACM symposium on Applied computing. pp. 2376–2380. ACM, New York, NY, USA (2008)
7. Domingos, P., Pazzani, M.J.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29(2-3), 103–130 (1997)
8. Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity search for web services. In: VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases. pp. 372–383. VLDB Endowment (2004)
9. Dumais, S.T., Platt, J.C., Hecherman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Gardarin, G., French, J.C., Pissinou, N., Makki, K., Bouganim, L. (eds.) CIKM. pp. 148–155. ACM (1998)
10. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI. pp. 1022–1029 (1993)
11. Fellbaum, C., editor. 1998. WordNet: An Electronic Database. MIT Press, Cambridge, MA.

12. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Dean, T. (ed.) IJCAI. pp. 668–673. Morgan Kaufmann (1999)
13. Günay, A., Yolum, P.: Structural and semantic similarity metrics for web service matchmaking. In: EC-Web. pp. 129–138 (2007)
14. Guo, H., Ivan, A.A., Akkiraju, R., Goodwin, R.: Learning ontologies to improve the quality of automatic web service matching. In: Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J. (eds.) WWW. pp. 1241–1242. ACM (2007)
15. Heß, A., Kushmerick, N.: Learning to attach semantic metadata to web services. In: International Semantic Web Conference. pp. 258–273 (2003)
16. Jones, S., Paynter, G.W.: Human evaluation of kea, an automatic keyphrasing system. In: JCDL. pp. 148–156. ACM (2001)
17. Jones, S., Paynter, G.W.: Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. JASIST 53(8), 653–677 (2002)
18. Lausen, H., Steinmetz, N.: Survey of current means to discover web services. Tech. rep., Semantic Technology Institute (STI) (August 2008)
19. Leung, C.H., Kan, W.K.: A statistical learning approach to automatic indexing of controlled index terms. JASIS 48(1), 55–66 (1997)
20. Lu, J., Yu, Y.: Web service search: Who, when, what, and how. In: WISE Workshops. pp. 284–295 (2007)
21. Ma, J., Zhang, Y., He, J.: Efficiently finding web services using a clustering semantic approach. In: CSSSIA '08: Proceedings of the 2008 international workshop on Context enabled source and service selection, integration and adaptation. pp. 1–8. ACM, New York, NY, USA (2008)
22. Peng, D., Huang, S., Wang, X., Zhou, A.: Management and retrieval of web services based on formal concept analysis. In: Proceedings of the The Fifth International Conference on Computer and Information Technology (CIT'05). pp. 269–275. IEEE Computer Society (2005)
23. Platzer, C., Dustdar, S.: A vector space search engine for web services. In: Third IEEE European Conference on Web Services, 2005. ECOWS 2005. pp. 62–71 (2005), http://dx.doi.org/10.1109/ECOWS.2005.5
24. Sabou, M., Wroe, C., Goble, C.A., Mishne, G.: Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In: Ellis, A., Hagino, T. (eds.) WWW. pp. 190–198. ACM (2005)
25. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. vol. 12. Manchester, UK (1994)
26. Turney, P.D.: Learning algorithms for keyphrase extraction. Inf. Retr. 2(4), 303–336 (2000)
27. Turney, P.D.: Coherent keyphrase extraction via web mining. In: Gottlob, G., Walsh, T. (eds.) IJCAI. pp. 434–442. Morgan Kaufmann (2003)
28. Wang, Y., Stroulia, E.: Semantic structure matching for assessing web service similarity. In: 1st International Conference on Service Oriented Computing (ICSOC03. pp. 194–207. Springer-Verlag (2003)
29. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)