Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Giuseppe Rizzo [a,*], Federico Tomassetti [b], Antonio Vetro' [b], Luca Ardito [b], Marco Torchiano [b], Maurizio Morisio [b] and Raphaël Troncy [a]

[a] *EURECOM, Sophia Antipolis, France*
E-mail: [giuseppe.rizzo, raphael.troncy]@eurecom.fr
[b] *Politecnico di Torino, Turin, Italy*
E-mail: [federico.tomassetti, antonio.vetro, luca.ardito, marco.torchiano, maurizio.morisio]@polito.it

Abstract. A Systematic Literature Review (SLR) identifies, evaluates and synthesizes the literature available for a given topic. This generally requires a significant human workload and has subjectivity bias that could affect the results of such a review. Automated document classification can be a valuable tool for recommending the selection of primary studies. In this paper, we propose an automated pre-selection approach based on text mining and semantic enrichment techniques. Each document candidate is firstly processed by a named entity extractor. The DBpedia URIs coming from the entity linking process are used as external sources of information. Our system collects the bag of words of those sources and it adds them to the initial document. A Multinomial Naive Bayes classifier discriminates whether the enriched document belongs to the positive example set or not. We used an existing manually performed SLR as golden dataset. We trained our system with different configurations of relevant papers and we assessed the goodness of our approach using a statistical approach. Results show a reduction of the manual workload of 18% that a human researcher has to spend. As baseline, we compared the enriched approach with one based on a normal Multinomial Naive Bayes classifier. The improvements range from 2.5% to 5% depending on the dimension of the trained model.

Keywords: Systematic Literature Review, Named Entity, Semantic analysis, Knowledge extraction, Automatic classification

1. Introduction

A Systematic Literature Review (SRL) is a mean to identify, evaluate and interpret all available interesting research to a particular topic area or phenomenon of interest. It has to be performed according to a predefined protocol describing how primary studies are selected and categorized, reducing as much as possible subjectivity bias. Depending on the research field where it is applied, the protocol changes. In this paper, we focus on a SLR applied to the field of Software Engineering, where the protocol can be summarized by the following steps [11]: (i) identification of research, (ii) selection of primary studies, (iii) study quality assessment, (iv) data extraction and monitoring progress, (v) data synthesis. The first step defines the search space, i.e. the set of document in which researchers select papers. A small sample set of relevant documents is used to define the search space. The second step identifies and analyses all possible useful studies

[*] Corresponding author. E-mail: giuseppe.rizzo@eurecom.fr.

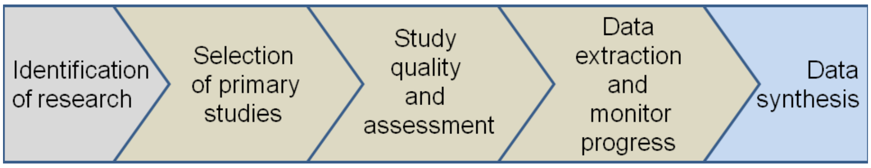
Identification of research
Selection of primary studies
Study quality and assessment
Data extraction and monitor progress
Data synthesis

Fig. 1. The big picture about the Semantic Literature Review process

among the papers which are contained in the search space that can help to answer some research questions. In the third step, an assessment about the quality of the studies collected is performed, while in the fourth step, the data extraction forms according to the review under evaluation is delivered. The last step delivers the data synthesis methods. Although these steps seem to be sequential, it is worth considering them as iterative steps and, therefore, the outputs may evolve according to the evolving topics. Figure 1 summarizes the five steps involved in the SLR process. The entire process is supervised and guided by researchers who summarize all existing information about some phenomenon in a thorough and, potentially, unbiased manner. The final goal is to draw more general conclusions about some phenomenon derived from individual studies, or as a prelude to further research activities. A SLR has a crucial importance in all research fields but it is extremely time-consuming, requiring an important human workload which is costly. The objective of this paper is to reduce the human workload in an SLR, semi-automating the selection of primary studies (i.e. the second step of the SLR process). This depends on the dimension of the search space. The larger the search space is and the more effective our proposed approach will be. Our method focuses on a filter strategy resorting on semantic enrichment and text mining techniques to reduce the number of papers that researchers, who perform a SLR, should read. We use a text classifier to filter potentially interesting documents within the search space. The classifier produces a reduced set that shall contain a higher interesting document percentage than the initial set. Afterwards, this reduced set is examined manually by researchers. In this way, we reduce the workload required to all researchers, limiting the human error rate. This phenomenon usually occurs when a set is sparse and searching through it requires more efforts than in a clean set, where the noise is smaller.

First, we are interested in investigating if the improved selection process allows reducing the manual work needed for the selection of studies while preserving completeness, and if yes, in quantifying the work saved (RQ1). Subsequently, we aim to assess the contribution of the semantic enrichment mechanism. We compare results achieved with a Multinomial Naive Bayes classifier in the case of analysis of enriched documents (obtained with the semantic enrichment) and in the case of examination of original papers (not-enriched ones) (RQ2). We address therefore the following research questions:

1. Does the automatic selection process based on the Multinomial Naive Bayes classifier and semantic enrichment (enriched process) reduce the amount of manual work of a SLR with respect to the original process?
2. Does the automatic selection process based on Multinomial Naive Bayes classifier and semantic enrichment (enriched process) reduce the amount of manual work of the alternative version of the process with only Multinomial Naive Bayes classifier (not-enriched process)? In other words, we aim to validate the idea behind the use of enriched papers as test samples instead of using original papers as test samples.

The work presented in this paper is based on a previous system presented by the same authors [18]. The following improvements are proposed: while previously the automatic classification was planned to fully automate the entire selection process step, in this paper, we propose a semi-supervised approach. This is because papers selected by the automatic classifiers could be immediately discarded by a human researcher just looking at the title and the abstract and do not need necessarily to be fully read. In addition, we perform an evaluation on a much larger dataset, extending the ground truth from the previous 111 papers to the current 2215 papers (almost 20 times larger). Finally, we propose an exhaustive evaluation based on the mentioned research questions.

The remainder of this paper is organized as follows. Section 2 details the step that focuses on the selection of primary studies and Section 3 presents our approach to improve this step. Section 4 describes the use case we use to validate our approach. In Section 5, we report and discuss the results we obtained. Section 6 compares our approach with the state of the art in the SLR field. Finally, we give our conclusions and outline future work in Section 7.

2. Selection of primary studies

In this section, we detail the selection step of the SLR process (Figure 1) analyzing its strengths and weaknesses according to the guidelines described in [11]. This step takes as input the set of primary studies W gathered from the universe of all scientific papers in the domain of interest of the review. W represents the result set of the first step of the process and it is obtained as the output of the search process performed by human beings using keywords on dedicated sources. For instance W could be composed by all papers published by a given set of journals or by all papers that a digital library provided as result of the search with keywords. The selection of primary studies is divided in two sub-steps: the former operates a selection based on reading titles and abstracts (*first selection*), the latter is the decision based on the full text human analysis (*second selection*). Both steps are basically affected by the following choice criteria: does it fit the research field? We define C (*candidate studies*) the set of studies that successfully passed the first selection and are eligible to be processed by researchers in the second selection step. It has the goal to split C in I (*included studies*) and E (*excluded studies*) where those sets are:

- I is the set of studies $\in C$ that successfully passed the second manual selection and will contribute to the systematic review. The following relation holds: $I \subseteq C$.
- E is the set of studies $\in C$ that did not pass the second manual selection and will *not* contribute to the systematic review and synthesis. Hence, $E \subseteq C$ and $E \cap I = \oslash$.

Figure 2 illustrates the selection of primary studies step. As introduced in the previous section, the selection of primary studies is performed by human beings who usually apply selection criteria which may be potentially biased by the level of knowledge about the topic that the reader has. A semi-supervised approach aims to reduce this potential bias.

3. Approach

The proposed approach relies on text mining techniques and semantic enrichment to reduce the set of interesting papers a researcher has to evaluate. The approach consists of a semi-supervised iterative process built on top of the following assumption: $W \neq \oslash$ (as a result of the applied search strategy) and $I \neq \oslash$ at the beginning (the set of relevant documents already known when the systematic review starts is not empty). The output of this approach is the set of most interesting papers W' gathered from a larger set of unread papers W.

3.1. I_0 construction

The initial set of sources contained in I is named I_0 and it is composed of primary studies already classified as relevant for the review: this is the first step of our process and it is needed to start the iterative part of the algorithm. I_0 can be built in two different ways. The first way is to ask researchers to use their previous knowledge indicating the most well known and fundamental papers in the field of interest. This strategy considers that often, systematic reviews are undertaken by experts in the field. The second way is to explore a portion of the search space using the basic process, e.g. searching on digital libraries or selecting the issues of (a) given journal(s). This portion is marked as I_0 and the enriched process is used to explore the remaining search space.

3.2. Model building

The second step of our approach consists in computing automatically a model M from I_0. The idea is to build a bag of words (BOW) model starting from the primary studies in I_0. For each study, we considered the words from the abstract and introduction. According to [4] terms that appear at the beginning and at the end of a document (such as title, abstract, introduction and conclusion) are more significant. We empirically assessed that using a reduced set of words, coming only from abstract and introduction, provides the same results of considering the extended set of words (i.e. set of words coming from the title, abstract, introduction and conclusion). The explanation is that the semantic enrichment approach compensates a reduced cardinality of the BOW through linking external sources and gathering from them textual data. Finally, we perform stop words elimination and stemming process, using

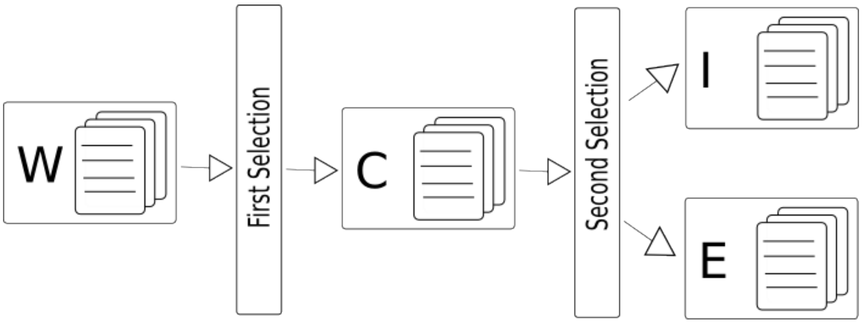

Fig. 2. Selection of primary studies in a Systematic Literature Review

the Porter algorithm [17]. The model built is used to train a Multinomial Naive Bayes classifier which computes the weight for each word according to the TF-IDF normalized approach [10].

3.3. Semantic enrichment

We define w_i a document composed by the BOW collected from the abstract and the introduction of one paper $\in W$. Each w_i is processed to get a set of named entities N which summarizes w_i. Formally, a named entity is a name of a person or an organization, a location, a brand, a product, a numeric expression including time, date, money and percent found in a sentence [6]. Basically, it is an information unit described by a set of classes (e.g. person, location, organization) and potentially univocally defined by external information (entity linking). If applied to the Linked Open Data (LOD) cloud [5], this operation is called disambiguation process [19] and it allows to point to resources published according to the Tim Berners-Lee's rules [1]. In our approach, we select information from DBpedia [2], the RDF structured information extracted from Wikipedia infoboxes. The encyclopedic nature of this dataset is appropriate to enrich the content of each w_i.

Once we have extracted the set of named entities N, we link each $n_i \in N$ to the corresponding DBpedia resource (when it is available). The extraction of named entities is performed using OpenCalais[1]. OpenCalais provides a classification for each named entity and suggests a URI of an external source where the information is disambiguated. Relying on it, we point to a DBpedia resource defined by the *owl:sameAs* property. If it is not available, we look for entities (i.e. subjects) in DBpedia which have the same labels extracted by OpenCalais. Once the resource is found, then we collect all words contained in the description field (`dbpedia-owl:abstract` property). The abstract property is one of the descriptive property which is consistent in the entire DBpedia dataset. After collecting these descriptions, we add them to the bag of words natively taken by the document w_i. We call it the enrichment process and the resulting document is defined as $w{+}_i$. Finally, it is compared with the trained model M using a Naive Bayes classifier which is described below.

3.4. Classification

We used a Multinomial Naive Bayes (MNB) classifier and we implement the TF-IDF weight normalization. According to [10], this implementation outperforms the CNB used by [12]. We use the classifier to compare $w{+}_i$ with the model M and we determine whether the conditional probability that $w{+}_i$ belongs to I is significant or not. We assume that all papers that do not belong to I, belong to E adopting the Boolean algebra. The comparison is done for each $w{+}_i \in W$: papers with $P[w{+}_i \in I] \geq threshold$ are moved to W' and they are manually analyzed by researchers. Finally, all the papers whose $P[w{+}_i \in I] < threshold$ remain in W.

3.5. Iteration

The papers with $P[w{+}_i \in I] \geq threshold$ are moved to W' to be manually processed, whilst the remaining ones still remain in W. It is likely that some of the papers moved in W' will pass the manual selection and will go to I, while the others will go to E. When I is modified, M becomes obsolete and it is necessary to re-build the model and repeat the classification step for all papers $w{+}_i \in W$. Again, if

[1] `http://www.opencalais.com`

$P[w+_i \in I] \geq threshold$, $w+_i$ is moved to W' to be manually analyzed. If any $w+_i$ goes to W', i.e. $W' = \oslash$ after a classification, the iteration stops. Papers that remain in W after the last iteration are finally discarded and not considered by researchers. The exclusion of these papers represents the reduction in workload for the human researchers. At each iteration, the model will be progressively tailored to the domain of interest, allowing to refine the selection of primary studies.

Algorithm 1 Enriched selection process algorithm

```
Define $I_0$
Init I with $I_0$
repeat
    /* automatic recommendation of primary studies */
    Train classifier with $I$
    Extract model $M$
    for all $w_i$ in $W$ do
        Enrich $w_i$ obtaining $w+_i$
        Compare $w+_i$ with model $M$:
        if P[$w+_i$ in $I$] $\geq threshold$ then
            move $w_i$ to $W'$
        end if
    end for
    /* first selection */
    for all $w'_i \in W'$ do
        Manually read title and abstract ($w'_i \in I$ ) ? move $w'_i$ to $C$ : discard $w'_i$
    end for
    /* second selection */
    for all $c_i \in C$ do
        Manually read full paper ($c_i \in I$ ) ? move $c_i$ to $I$ : move $c_i$ to $E$
    end for
until $C \neq \oslash$
Discard $\forall\, w_i \in W$
```

We provide in Algorithm 1 the synopsis of the whole study selection process proposed in this paper and in Figure 3 its complementary graphical representation. Comparing this picture with Figure 2 that represents the selection process provided by the guidelines [11], we observe that the original process is not changed, but we have added a selection of primary studies that recommends papers similar to the model at each iteration. We also reported in Figure 3 the steps of the new process described in subsections 3.1 to 3.4: the use of a model of bag of words (b) derived from I_0 or I (a), the enrichment of papers through semantic enrichment (c) and the comparison of the model M with the studies through a Multinomial Naive Bayes classifier (d).

4. Experiment

We have developed the Semantic Systematic Review tool and we have released the source code at http://sourceforge.net/projects/semreview

The tool allows to load an already performed SLR from which are already known both the set of interesting papers and the set of not-interesting ones. This enables to run experiment and assess the effectiveness of our approach. The tool creates the initially set of relevant papers I_0 (papers which belong to the I set) selecting randomly a sub-set of the interesting papers defined by the SLR. Doing it, the tool simulates the operation performed by human researchers at the beginning of the SLR. The other interesting papers, together with the not-interesting ones, end in the W. This set is used for assessing the performance of the approach. From I_0, the tool extracts the corresponding BOW and initializes the model M. Then, for all the papers in W, the tool performs automatically the recommendation of the primary studies (the second step in the SLR process) implementing the approach described in Section 3. Finally, the tool reports the performance of the approach using as ground truth the SLR taken as reference. The performance is measured as the amount of the saved manual work. The baseline in the experiment is given by the semi-supervised automatic approach without the semantic enrichment mechanism.

4.1. Golden dataset

As a case study we selected a SLR on Software Cost Estimation done by [9] and we limit the ground truth to all the papers mentioned in the SLR coming from the IEEE Transactions on Software Engineering (IEEE TSE) journal. They cover a timeframe ranging from 1977 to April 2004. Unfortunately, we had to exclude the first volume of IEEE TSE because it is not accessible from the IEEEXplore portal[2]. The resulting set contains 2215 candidates, all of them evaluated from the SRL taken as reference. The original SLR contains 51 interesting papers. However, only 50 of them are actually present in the set of the candidates available from the IEEEXplore, the missing one having been published in the first volume of IEEE TSE. Our golden dataset is therefore composed of 2215 papers, and 50 of them belong to the I set. The others are considered as not-interesting papers, i.e. they do not pass the selection criteria defined at the beginning of the performed study and they belong to the E set.

[2] http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=32

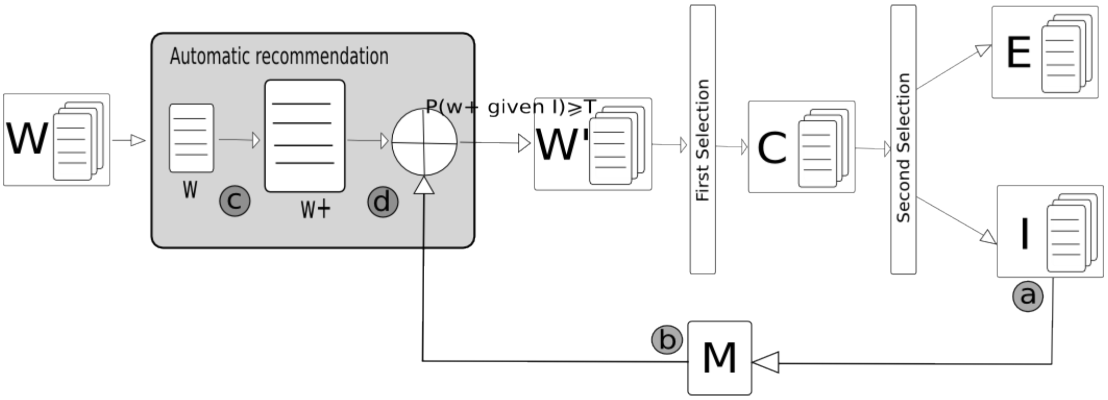

Fig. 3. The enriched study selection process and its principal steps: model extraction (b) after I is built (a), enrichment of papers through semantic enrichment (c) and comparison with the model through a Multinomial Naive Bayes classifier (d).

4.2. Variable selection

The main outcome under measurement is the manual work, consisting of reading primary studies either entirely or only title and abstract, to select the interesting ones for the subject of the SLR. We measure the manual work as the number of papers that are read assuming the number as a proxy for the actual time that would be spent reading the articles. The minimum manual work ideally required is the total number of interesting papers. However, this minimum could reasonably never be reached in SLR. Indeed, the relation $I \subset W$ holds, where I is the set of relevant papers and W is the set of containing papers defined by the search criterion. This choice is motivated by the fact that the SLR, selected as subject of the case study, does not report neither the time spent for papers selection nor which papers were read entirely and which partially (only title and abstract). As a consequence, we define the following two metrics:

mw is the manual work. More specifically mw_O is the manual work performed in the original SLR, i.e. manually selecting and reading all papers, mw_{NE} is the manual work obtained applying the selection based on the Multinomial Naive Bayes classifier using original papers (not-enriched process), mw_E is the manual work obtained applying the selection based on the Multinomial Naive Bayes classifier using enriched papers (enriched process).

t is the applied task. Three levels are possible: manual, not-enriched, enriched.

4.3. Hypothesis formulation

The last step of the design is the hypothesis formulation. We formulate a pair of null and alternative hypothesis for each of the two research questions. Goal of the experiment is to reject the null hypothesis H_0 monitoring the `p-value` [8]. In other words, we discard the null hypothesis and we validate the alternative one H_A if the probability to reject the H_0 is lower than the 0.001. Moreover, it tells that when choosing the alternative hypothesis H_A, the probability to commit an error is lower than 0.001.

1. $H1_0 : mw_O \leq mw_E$, recall= 0.95
   $H1_A : mw_O > mw_E$, recall= 0.95

2. $H2_0 : mw_{NE} \leq mw_E$, recall=0.95
   $H2_A : mw_{NE} > mw_E$, recall=0.95

4.4. Parameter configuration

We decided to assess the validity of our process with different size of I_0 ranging between 1 and 5. In order to limit the bias introduced by a particular configuration of selected papers, we built 30 different I_0 sets per each dimension choosing them randomly among 50 relevant papers. We used each generated I_0 to kick-off the two variants of the process: enriched and not-enriched. Moreover, we replicated the experiment varying the classification threshold between 0 and 1 with steps of 0.01. The classifier threshold represents the posterior probability for a sample to belong to the I (interesting set). Overall, we executed the complete algorithm 30,300 times = 5 (number of I_0 sizes) x 30 (number of I_0 sets for each size) x 2 (variants of the algorithm) x 101 (thresholds).

A preliminary step consisted to define the best classifier threshold T which maximizes the recall for the two variants. According to [4], we decided to aim at a recall of 95%. Although this recall value is a strong constraint, we adopted it for limiting as much as possible the elimination of interesting papers. Hence, we have empirically observed that the threshold is also adequate for other datasets in the same area (SLR of software engineering). In Table 1, we report the distribution of the maximum classifier threshold which permits to obtain the target recall using the different I_0 sets. We chose the maximum threshold because is the one which minimizes the workload while it still satisfies the requirement of a recall equal to or greater than 95%. We select the median values to set the classifier, that means 0.22 for the enriched process and 0.17 for the not enriched one.

4.5. Analysis methodology

The goal of data analysis is to apply proper statistical tests to reject the null hypotheses we formulated. Since the values are not normally distributed (according to the Shapiro test), we adopt a non parametric test. In particular, we select the Mann-Whitney test [7] that compares the medians of the vectors of mw. To do that, we considered all papers extracted from the dataset except those papers used to build the I_0.

5. Results and Discussion

Figure 4 shows the comparison distributions for different settings of I_0 according to the two different type of recommendation approaches proposed: enriched process or not-enriched process. On the y-axis, the workload needed for a human being after both processes (enriched E and not-enriched NE) is reported. On the x-axis, we indicate the number of papers used for training the I_0 set and the process used (e.g. 1.E means an I_0 composed of 1 paper and the process has been performed using the enrichment mechanism). We observe a reduction of the workload in both approaches. Comparing the semantic enrichment with the baseline, we observe a greater reduction of the workload. This increment ranges from 2.5% to 5% for all I_0 settings, except for the I_0 composed of 1 paper (1.E in Figure 4) where the increment is lower then 1% with respect to the not-enriched (e.g. 1.NE in Figure 4).

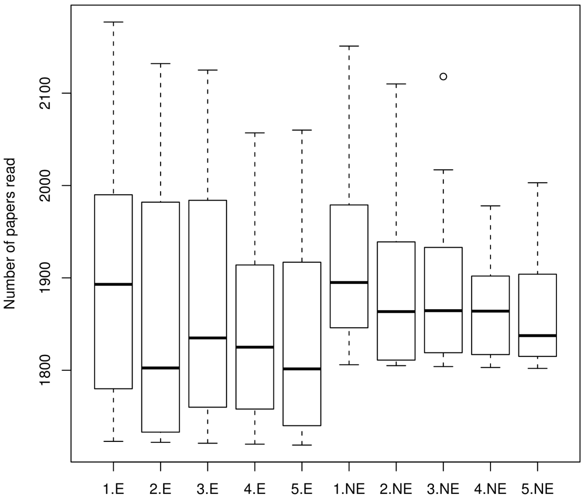

Fig. 4. Number of papers to read for different I_0 sizes and tasks applied: E (with enrichment) and NE (without).

We present below the results according to the two research questions addressed in this paper (see Section 1): evaluating whether the semantic automatic process classification reduce the amount of work of a SLR or not (RQ1) and evaluating if the semantic enrichment increases the performance of the simple classification process (RQ2).

5.1. RQ1: Reduction of the Human Workload

The results from the Mann-Whitney test are shown in Table 2. The table reports the I_0 size (column 1), the manual work in the original SLR process (column 2), the manual work obtained with our enriched process (column 3), the estimated percentage of manual work to be performed with our enriched approach with respect to the total work required using the common approach (column 4) and the `p-value` obtained from the Mann-Whitney test. The `p-value` for all the configurations indicates that the null hypothesis can be rejected and we assume the alternative which motivates the choice to use the semantic enrichment approach. In addition, we notice that the workload reduction increases as the size of I_0.

5.2. RQ2: Assessing the Performance of the Enrichment Process

We used the Mann-Whitney test to reject the null hypothesis by which we state that $mw_{NE} \leq mw_E$. Ta-

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
not-enriched	0.11700	0.1700	0.1700	0.1729	0.1775	0.1900
enriched	0.2100	0.2100	0.2200	0.2201	0.2200	0.2600

Table 1

Analysis of the best classifier threshold for both enriched and not-enriched process across different I_0 sets. The first and last column show the minimum and maximum values, second and fifth columns respectively the first and third quartile of the distribution, then mid columns show median and the mean of it

	Workload		Manual workload vs enriched workload	
$\lvert I_0\rvert$	mw_O	mw_E	median	$p - value$
1	2214	1897.567	**85%**	< 0.001
2	2213	1864.367	**84%**	< 0.001
3	2212	1863.833	**84%**	< 0.001
4	2211	1843.133	**83%**	< 0.001
5	2210	1829.1	**82%**	< 0.001

Table 2

For each I_0 configuration, we first compare the workload required to a human being in the original SLR and the workload mean if our process is performed. To verify the goodness of our process, we compute the Mann-Whitney test and we reject the hypothesis $mw_O \leq mw_E$ with a recall = 0.95.

$\lvert I_0\rvert$	workload median pairwise difference	$p - value$
1	26.67	0.0192
2	66.00	0.0073
3	40.83	0.0090
4	33.00	0.0083
5	**49.99**	0.0009

Table 3

For each I_0 configuration, we performed the Mann-Whitney test, evaluating median pairwise difference and `p-value` to estimate the minimum workload using both process: enriched and not-enriched. As for RQ1, the minimum recall is 0.95.

ble 3 reports the I_0 size (column 1), the estimated difference of manual workload between the two processes (column 2), and the `p-value` of Mann-Whitney test (column 3). While we can observe that the enriched process requires less workload for every size of I_0, we can affirm it with $p < 0.001$ just when the size of I_0 is 5.

5.3. Discussion

The results show that our approach actually reduces the human workload to perform a SLR, while aiming to maintain a high level of completeness. Indeed, by limiting the recall to 95%, we adhere to the state of the art in the automation of SLR field maintaining its high quality. However, relying only on positive papers, this approach introduces one more configuration step for defining the threshold. The threshold can changes according to the field of the SLR. In our test, we empirically observed that the probability threshold is almost consistent in different test scenarios. For this reason, we consider it as a baseline value for further researches.

In addition, we observed that the enriched process performs better than the variant without enrichment (almost 2.5% when I_0 is composed of 5 papers). There are still two shortcomings: `i)` the extracted entities from OpenCalais sometimes point to URIs which do not adhere to the Linked Data principles, and/or do not contain sameAs links to DBpedia resources. We observe that the enrichment process fails in around 20% cases. The fallback strategy to rely on another interlnking step using the named entity labels and lookup in DBpedia does not entirely fill the gap, mainly because it evaluates the label similarity between the target entity (extracted from OpenCalais) and the entities stored in DBpedia without considering the context from which the name entity has been extracted (ambiguity problem). `ii)` a massive use of encyclopedic sources can bias the content of the enriched paper, penalizing words which do not appear often in the linked source but that are frequent in the initial document.

Differently from what we expected, the I_0 configuration does not affect the recall. Indeed, our results suggest that the number of papers in I_0 is not relevant. Its composition in terms of which papers are used to create it may play a more important role. For instance, let's consider an initialization of I_0 with papers that are not strictly related or if they represent just a niche of the research field, or if we select papers which are completely out of argument and they represent different meaning. While in the latter case, a wrong initialization affects all process and requires the initial set, in the former case the enrichment process enlarges I evading from the niche. Experiments show that subjective bias in the composition of I_0 is reduced when we use the semantic enrichment approach. While we do not have statistical evidence for that, I_0 size seems to play a role on workload reduction.

6. Related Work

The automatic text classification applied to a systematic review is more challenging than the typical classification task. This is basically due to the dynamic nature of a SLR which is a supervised and iterative process where the initial scope of the SLR often evolves during the review process. Numerous research efforts have been spent to reduce the human workload when a SLR is performed. We focus on two different types of studies: i) machine learning based, and ii) ontology based.

Cohen *et al.* proposed a first attempt to reduce the human workload in the SLR field [4]. They used the automatic classification to discard not-interesting papers from a set of them in fifteen different medical systematic literature reviews, each one considering the validity of a particular drug. Their classification model uses a reduced set of the features gathered from the paper such as author name, journal name, journal references, abstract, introduction and conclusion. The classification model is built using both positive and negative examples, where negative examples are selected from the pool of papers which do not adhere to the chosen SLR. Finally, this model is used to create a perceptron modified vector for each feature in the feature set. Negative examples bias the model. In order to limit this phenomenon, they introduced a perceptron learning adjustment just evaluating the false negatives and false positives, monitoring them according to the False Negative Linear Rate (FNLR). A test article is classified by taking the scalar product of the document feature vector with the perceptron vector and comparing the output values. Considering a recall of 95%, the reduction of workload ranges from 0% to 68% according to the SLR they took under evaluation. Among the findings, they suggested that automatic classification may be useful to monitor regularly new relevant journal issues in order to identify interesting primary studies, easing the task to keep a SLR constantly updated. According to this result, it is crucial to consider the classification problem in the SLR field as a semi-supervised approach in which a human being supervises the inclusion or exclusion of possible relevant studies selected by the classifier. Another attempt to reduce the human workload in selecting relevant primary studies was performed by [12]. They proposed an approach mainly based on the Naive Bayes classifier with some optimizations which are based on the Complement Naive Bayes (CNB) [13]. The results they achieved outperform what detailed in [4], but using a different configuration parameters (they consider only title and abstract for each document instead of the large set of features considered by Cohen). Leveraging on Natural Language Processing techniques (NLP), Cohen *et al.* tackle the problem of paper handling once the review starts [3]. This is practically done to allow the reviewer to first analyze the documents which are labelled as potentially relevant documents, leaving at the end the evaluation for the remaining ones. They combined the approach of unigram and Medical Subject Headings (MeSH) to create the histogram of documents which potentially fits the scope of the review.

Ruttenberg *et al.* [16] proposed an hybrid approach for automating scientific literature search by means of data aggregation and text mining algorithm to make easy the search process. The key point of their work was to find a way to represent and share knowledge learned by human beings reading relevant papers, by means of an ontology. Through it, it was possible to combine outcomes of each single document and to represent it into a graph, which is mapped to the ontology. The first step of this process consists of identifying the key phrases of the document (outcomes). Then, key phrases are used to link different concepts in the graph. Following this process, concepts are linked together, obtaining a chain of relationships. This work is usually made by human beings, who are experts of the domain. Ideally, they are impartial but the authors assessed that the graph mapping is strongly affected by the expert subjectivity. Then, they proposed a mechanism based on text mining algorithms to be able to navigate and cluster inferences. This work represents the first attempt to introduce the concept of knowledge representation in a SLR and, among the findings, they stated that a pre-clustering and linking of documents limit the human subjectivity improving the overall result.

7. Conclusion and Future Work

In this paper, we presented a semantic enrichment recommendation of primary studies in a SLR. Resorting on text mining techniques and semantic enrichment, we improved the second step of the SLR process in order to filter the set of possible studies a researcher should read, automatically discarding the not relevant papers. Our approach has two main advantages: i) reduction of workload requested to classify sources and ii) reduction of subjectivity in the overall process. We tested our approach using a real SLR [9] which is used as ground truth. Keeping a recall of 95% (i.e. we expected to discard papers only when the system is at least 95% sure that the paper is out the scope) we gained a percentage of workload saved of 18% when I_0 is composed of 5 papers. In addition, we demonstrated that the enrichment process outperforms of 2.5% the basic classification process (i.e. the process which takes into account only original documents) with an I_0 composed of 5 papers. The automatic recommendation process without enrichment is used as baseline for validating the semantic enrichment approach.

As future work, we have planned to improve substantially the classification step, using besides positive examples also negative examples. We believe that using also negative examples the process may have a more accurate value of the plausible probability if a sample belongs to the interesting set. The first idea is to use some of the papers not included in the SLR for training negative examples. Although this may be intuitive, we may address the problem of a short distance from positives and negatives, due to the cross topics which these papers may report. A further evaluation of the distance among papers from different journal issues may give a better idea about the use of negative examples. Therefore a deep analysis of which studies may be considered as negative is needed. In addition, we have planned to extract one paper i at a time from the set of relevant papers I, and use the remaining papers $\in I$ to train the classifier and then evaluate if it recognizes i as similar to the others. In this way, the classifier is used to give a "second opinion" on the selection process, potentially reducing the number of researchers necessary to undertake this step.

In the presented approach, we rely on the MNB classifier. It is considered as the baseline for text classification, but its results are often comparable to the state of the art in text classification (such as SVM and Markov chain) [13]. We have planned to validate the use of the semantic enrichment with state of the art classifiers in order to understand also if we can improve the performance of the entire process. The experiments addressed an important weakness in the named entity extraction task. The disambiguation mechanism provided by OpenCalais often links each information unit with web resources in its authority domain and provides occasionally sameAs link to other LOD datasets (basically DBpedia). We have planned to use NERD [15,14] to have a better disambiguation process, leveraging on the aggregation of several named entity extractors and to perform an assessment on the costs, precision and recall of the named entity extraction step. Finally, the semantic enrichment mechanism has been validated using one SLRs. We have planned to validate it also using other SLRs especially coming from other field of research.

We believe that our approach could be adopted by scientific content providers such as journal portals, to index sources and to automatic classify and cluster the papers they publish. This approach may be used to propose a faceted view of sources queried by a user. The challenge will be to compute this operation in real-time for limiting human efforts.

Acknowledgments

The research leading to this paper was partially supported by the French National Agency under contracts ANR.11.EITS.006.01, "Open Innovation Platform for Semantic Media" (OpenSEM) and the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911).

References

[1] T. Berners-Lee. Linked data, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.

[3] A. M. Cohen. Optimizing feature representation for automated systematic review work prioritization. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, pages 121–125, 2008.

[4] A. M. Cohen, W. R. Hersh, K. Peterson, and P. Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association (JAMIA)*, 13(2):206–219, 2006.

[5] Richard Cyganiak and Anja Jentzsch. Linking Open Data cloud diagram. LOD Community (`http://lod-cloud.net/`), 2010.

[6] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: a brief history. In *16th International Conference on Computational linguistics (COLING'96)*, pages 466–471, 1996.

[7] Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. John Wiley and Sons, New York, 1973.

[8] Raymond Hubbard and R. Murray Lindsay. Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology*, 18(1):69–88, 2008.

[9] M. Jorgensen and M. Shepperd. A Systematic Review of Software Development Cost Estimation Studies. *IEEE Transactions on Software Engineering*, 33(1):33–53, 2007.

[10] Ashraf Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial Naive Bayes for Text Categorization Revisited. In *17th Australian joint conference on Advances in Artificial Intelligence (AI'05)*, volume 3339, pages 235–252, 2005.

[11] B. Kitchenham. Procedures for performing systematic reviews. Technical report, 2004.

[12] Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association (JAMIA)*, 17(4):446–453, 2010.

[13] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *20th International Conference on Machine Learning (ICML'03)*, pages 616–623, 2003.

[14] Giuseppe Rizzo and Raphaël Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for computational Linguistics (EACL'12)*, 2012.

[15] Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Bruemmer. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In *5th International Workshop on Linked Data on the Web (LDOW'12)*, Lyon, France, 2012.

[16] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in Bioinformatics*, 10(2):193–204, 2009.

[17] Karen Sparck Jones and Peter Willett, editors. *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., 1997.

[18] Federico Tomassetti, Giuseppe Rizzo, Antonio Vetro, Luca Ardito, Marco Torchiano, and Maurizo Morisio. Linked Data Approach for Selection Process Automation in Systematic Reviews. In *Evaluation and Assessment in Software Engineering (EASE'11)*, 2011.

[19] Yorick Wilks and Christopher Brewster. Natural Language Processing as a Foundation of the Semantic Web. *Foundations and Trends in Web Science*, 1(3):199–327, 2009.