Linked Data approach for selection process automation in Systematic Reviews

Federico Tomassetti, Giuseppe Rizzo, Antonio Vetro', Luca Ardito, Marco Torchiano and Maurizio Morisio
Dipartimento di Automatica ed Informatica, Politecnico di Torino
c.so Duca degli Abruzzi, 24, 10129, Torino, Italy
Email: [federico.tomassetti|giuseppe.rizzo|antonio.vetro|luca.ardito|marco.torchiano|maurizio.morisio]@polito.it

***Abstract*—Background: a systematic review identifies, evaluates and synthesizes the available literature on a given topic using scientific and repeatable methodologies. The significant workload required and the subjectivity bias could affect results.**

Aim: semi-automate the selection process to reduce the amount of manual work needed and the consequent subjectivity bias.

Method: extend and enrich the selection of primary studies using the existing technologies in the field of Linked Data and text mining. We define formally the selection process and we also develop a prototype that implements it. Finally, we conduct a case study that simulates the selection process of a systematic literature published in literature.

Results: the process presented in this paper could reduce the work load of 20% with respect to the work load needed in the fully manually selection, with a recall of 100%.

Conclusions: the extraction of knowledge from scientific studies through Linked Data and text mining techniques could be used in the selection phase of the systematic review process to reduce the work load and subjectivity bias.

I. Introduction

A systematic review is a literature review performed to answer a set of research questions and has to be performed according to a pre-defined protocol describing how primary sources are selected and categorized. It strives to produce an objective evaluation of findings available on a certain topic, reducing as much as possible subjectivity bias. A systematic review is composed by five steps (Kitchenham [2004]): (i) identification of research, (ii) selection of 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 inside which researchers may select papers. Then, every research document fallen out is not treated in the selection process. The second step represents an attempt to identify and analyze all possible useful studies to answer the research questions among the papers which are contained in the search space. The selection workload is proportional to the dimension of the search space, consequently a large one determines a great deal of work to be done manually. Moreover, being an operation performed manually, the human opinion might influence the outcome. Our approach focuses on improving the second step resorting on text mining techniques and Linked Data. We use a text classifier to filter potentially relevant documents within the search space. The classifier produces a reduced set that shall contain a higher relevant document than the initial set, in our intentions. That reduced set is examined by researchers for the final selection. In this way, we reduce the workload required to all researchers, limiting human error rate. This phenomenon usually occurs when a set is sparse and searching on that may require more fatigue than in a clean set, where the noise is smaller.

The remainder of this paper is organized as follows: a review of the current state of the art in Section 2, key ideas of our approach are introduced in Section 3, then we enroll classification process augmented with structured information for the selection process in Section 4 . Afterwards, we present a case study conducted to validate the process in Section 5, we discuss threats and benefits in Section 6, and finally conclusions and future works are in Section 7.

II. Related work

Use of automatic classification was already explored for systematic reviews in the medicine field. Cohen et al. [2006] experimented automatic classification in fifteen different systematic reviews, each one considering the validity of a particular drug. Their classification model used a reduced set of structured data gathered from author names, journal name and other journal references, abstract and introduction to make the classification model. Then they performed the automatic classification, in order to obtain workload reduction. After applying the search strategy, a large pool of primary sources was obtained. Normally this pool would be completely analyzed by researchers but Cohen et al. used automated classification to discard not relevant papers from this pool, reducing the number of papers that researchers have to analyze. Considering a recall of 95% they obtained a variable reduction of sources to manually consider. On the different systematic reviews it ranged from 0% to 68%. Moreover they suggested that automated classification could be useful to monitor regularly new relevant journals issues in order to identify relevant primary sources and pointing them out to interested researchers, easing the duty to keep a systematic review constantly updated. In our approach we use the automatic classification in order to reduce workload related to the selection process and, also, we consider the entire bag of words of the article instead of the reduced set.

Another important aspect to be considered is the subjectivity in the selection of studies. Peinemann et al. [2008] conducted a study on disagreement in primary study selection analyzing five different reviews performed on the same topic (negative

pressure wound therapy) and considering the same time interval. They observed different selections of primary studies and therefore conclusions. Disagreement in primary sources selection can be partially due to different selection criteria (e.g. the choice to include or exclude papers written in German) but also the subjectivity played a role.

Even though the guidelines proposed by Kitchenham [2004] are widely referred in systematic reviews in the software engineering field, some criticism emerged. For instance, Staples and Niazi [2007] adopted a two-step process in primary sources selection, discarding immediately those appearing irrelevant and considering carefully just a limited subset of all retrieved sources. Moreover selection of sources was performed by just one researcher and not by at least two as suggested in Kitchenham [2004]. It is important to note that deviations from guidelines are motivated by the need for workload reduction to make it viable in more situations and therefore more widely applicable. A systematic review related problem is the representation of concepts contained in papers. Ruttenberg et al. [2009] 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 topical papers, by means of an ontology. Using it, it was possible to combine outcomes of each single paper and to represent them into a graph, which is mapped to the ontology. So that, papers were read in order to highlight key phrases (outcomes); although this process was driven by domain experts who, usually, are impartial, the tricky point was the subjectivity related. Key phrases were used to link different concept in the graph. Following this process, many concepts were linked between, obtaining chains of relationships. Moreover, authors proposed text mining algorithms able to navigate and cluster inferences.

We start from the idea of semantic representation of knowledge, but we use it for linking topical information available in papers to DBpedia Bizer et al. [2009], a well-know people heritage knowledge, by means of the Linked Data principles[1]. According to this process, we enrich the data space of articles with information useful to identify concepts and we use this model for the classification step. Our approach permits to augment the classification process by means of Linked Data approach and, moreover, reduces subjectivity related to the selection process.

III. Study Selection Process

The first step in the approach presented by Kitchenham [2004] is the identification of research. The aim of this phase is to identify a subset of articles, W (the working area gathered from the universe of all scientific papers), in the domain of interest applying the defined search strategy. 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 following step is the selection process which operates on W to obtain the primary sources to consider in the review. This process is performed by researchers and it 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*). 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 one. This second sub-step, in fact, 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 didn't 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 1 represents the selection process and the sets.

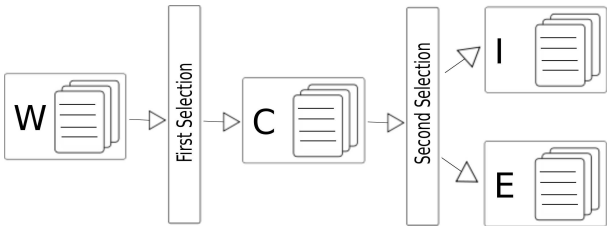

Fig. 1. Study selection process in systematic reviews (according to Kitchenham guidelines) represented through sets, selections and their relationships

IV. Enriched Study Selection Process

Our idea is to extend and enrich the basic process for the selection of primary studies using the existing technologies in the field of Semantic Web and text mining techniques, in the context of the Linked Data approach. The process we describe here is a supervised iterative process built on the top of the following assumption: $W \neq \oslash$ (as a result of the applied search strategy) and $I \neq \oslash$ at the begin (some relevant papers are already known when the systematic review starts).

A. I_0 Construction

The initial set of sources contained in I is named I_0 and it is composed by primary sources already classified as relevant for the systematic 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 researchers 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

[1]http://www.w3.org/DesignIssues/LinkedData.html

I_0 and the enriched process is used to explore the remaining search space.

B. Model Building

The second step of our approach consists in building a model M from I_0. The idea is to build a bag of words model starting from the primary studies in I_0. The bag of words model is a representation of the text as an unordered collection of words, holding their combined appearance frequency, disregarding grammar and word order (e.g. the sequence of words "data mining" has the same probability as "mining data"). For each primary study, we will consider only title, abstract, introduction and conclusion: according to Cohen et al. [2006], terms that tend to appear at the start or at the end of a document are the most significant ones. Then, we perform stop words elimination and stemming process, using the Porter algorithm, called Snowball[2]. The model so built represents our training set, used to make comparisons with candidates. Results of comparison are used as a classification measure and they suggest if a candidate belongs to the model.

C. Linked Data enrichment of papers

As described above, we adopt Semantic Web technologies to provide unambiguous space for identifying concepts highlighted in papers. In particular, this process uses the Linked Data approach which, therefore, addresses the exploitation of the Web as a platform for data and information integration. The main actor of this process is DBpedia[3], a Resource Description Framework (RDF)[4] repository where information stored in Wikipedia is represented as structured data. This repository works as a look-up system of resources. We define w_i a paper $\in W$: each w_i is processed to get a set of key-phrases K which describes w_i. This operation can be done using one of the common key-phrase/keyword extraction tools available. After that, we link each $k_i \in K$ to the correspondent DBpedia resource (when it is available). Results are mapped in a graph whose root is the requested resource, edges are the predicates which point to the objects (literals or URIs).

Then, we gather all words from statements of this property and we add these to the bag of words natively taken by the paper w_i. We call it enriching process and the resulting paper is named $w{+}_i$. Finally, it is compared with the trained model, M, by means of the Naive Bayes classifier, which is described below.

D. Naive Bayes Model Classification

We propose for the classification step a well-know approach in the text mining field: Naive Bayes classifier. This classifier is a very mature tool for classifying text documents, in particular it has been largely adopted in the e-mails classification in "spam and ham".

We use the Naive Bayesian classifier to compare $w{+}_i$ with the model learned and we determine whether the conditional probability that $w{+}_i$ belongs to I (from which M derives) is significant. We assume that all papers that do not belong to I, belong to E adopting the Boolean algebra. For this reason we only consider whether a document $w{+}_i$ belongs to I:

$$P(I|w{+}_i) = \frac{P(I) * P(w{+}_i|I)}{P(w{+}_i)}. \quad (1)$$

As described before, we build bag of words (terms), t_i from $w{+}_i$ and we consider each single word as a independent event $w{+}_i = t_1, t_2, t_3, ..., t_n$. Then, we obtain:

$$P(I|t_1, ..., t_n) = \frac{P(I)P(t_1, ..., t_n|I)}{P(t_1, ..., t_n)} \quad (2)$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on I and the values of the features t_i are given, so that the denominator is effectively constant. Then, we apply the assumption of the statistic independence of words in order to reduce the model complexity although we preserve good performances as presented in Schneider [2003]. As consequence the formula above becomes:

$$P(I|t_1, ..., t_n) = P(I) \prod_i P(t_i|I) \quad (3)$$

The comparison is done for each $w{+}_i \in W$: papers with $P[w{+}_i \in I] \geq threshold$ are moved to C, that means they automatically passed the first selection (that in the basic process is done manually) and become candidate studies. We select a threshold value based on the required recall. Considering the high recall we obtain a low precision and, consequently, a lower precision means a lower workload reduction.

The next step in our enriched process is identical to the second selection of the basic process: papers in C (candidates papers) will be manually read and included (go to I) or excluded (go to E). The decision impacts to the trained set, requiring a rebuilding of the model.

E. Iteration

So far the enriched process proposed is supervised ($I_0 \neq \oslash$ and the second selection is still manual) and multistage. We now add another characteristic: iterative. As described above, the papers with $P[w{+}_i \in I] \geq threshold$ are moved to C to be manually processed, whilst the remaining ones still stay in W. Likely some of the papers moved in C will also pass the second (manual) selection and will go to I, while the others will go to E. Then, whether I is modified, then M becomes obsolete and it is necessary to re-build it. We stop to iterate when $C = \oslash$. The papers that remain in W after the last iteration are discarded. At each iteration the model will be progressively tailored to the domain of interest, permitting to refine the selection process.

We provide below the synopsis of the whole study selection process proposed in the paper, in the form of the Algorithm 1, together with a complementary graphical representation (Figure 2). Comparing this picture with Figure 1, that represents the selection process provided by guidelines Kitchenham

[2]http://snowball.tartarus.org/texts/introduction.html
[3]http://dbpedia.org/About
[4]http://www.w3.org/RDF/

Algorithm 1 Enriched selection process algorithm

```
Define I_0
Init I with I_0
repeat
    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] ≥ threshold then
            move w_i to C
        end if
    end for
    for all c_i ∈ C do
        Manually (c_i ∈ I ) ? move c_i to I : move c_i to E
    end for
until C ≠ ∅
Discard ∀ w_i ∈ W
```

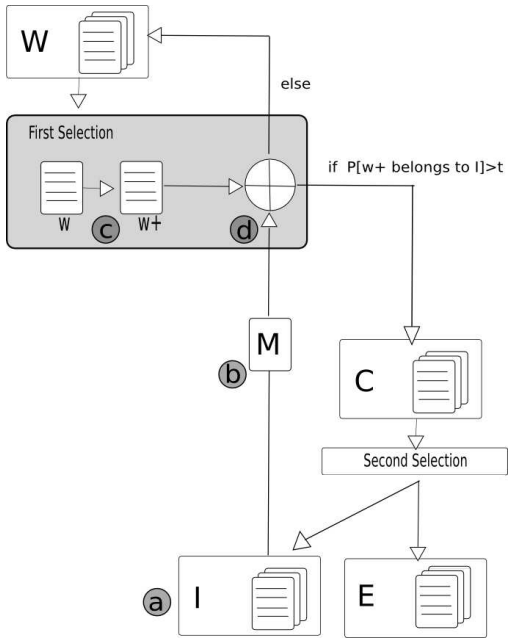

Fig. 2. The enriched study selection process and its principal steps: model extraction (b) after I is built (a), enrichment of papers through linked data (c) and comparison with the model through Bayesian classifier (d).

[2004], we observe the transformation of the first manual selection in a fully automatic selection. We also reported in Figure 2 the steps of the new process described in sub sections from IV-A to IV-D: the use of a model of bag of words (b) derived from I_0 or I (a), the enrichment of papers through linked data (c) and the comparison with the model M by means of the Bayesian classifier (d). For the sake of simplicity, we didn't represent the iteration (subsection IV-E).

V. Case study

We implemented a Java prototype of the algorithm and performed a case study to evaluate how the supervised process could work in a real systematic review. We selected as a reference a systematic review on Software Cost Estimation done by Jorgensen and Shepperd [2007]. The authors firstly selected a list of journals of interests, then they examined the title and the abstract of all the papers appeared in the issues of these journals in order to select which papers download. Finally they carefully read the downloaded primary studies to find which were relevant for the review. Our idea is to simulate a portion of their manual selection and check if our semi-automatic process could reduce the human workload without loosing any interesting paper. The case study design is the following: we select from Jorgensen and Shepperd [2007] the journal containing the highest number of relevant papers, i.e. IEEE Transactions on Software Engineering (TSE), then we search on IEEEXplore all papers in TSE using the search term *Software Cost Estimation* and with publication date from 1996 to 2004 (the year of the most recent paper of TSE in Jorgensen and Shepperd [2007]): we obtain 135 primary studies, some of which were immediately discarded because they were just indexes and not proper papers. The papers available for the case study are 111, 24 of which were included in the systematic review; we considered them as the relevant paper set. Afterwards, we create the bag of words model for each downloaded paper and we initialize the sets W and I_0 in the following way: I_0 contains 5 papers from the 24 relevant, whilst the 106 remaining studies, 19 of which are relevant, are in W. After the initialization is concluded, our prototype is able to perform automatically the remaining steps: the extraction of keywords (i.e. the social tags identified by the OpenCalais web service[5]), the enrichment of the model linking keywords to DBpedia abstracts, the classification of the enriched paper with the model of I with the Naive Bayes classifier, and finally the simulation of the second manual selection, that is automatized because we know *a priori* which papers are relevant.

The process is repeated using all possible thresholds of the Naive Bayes classifier (from 0.01 to 0.99) and replicated eliminating the enrichment process, i.e. using the simple bag of words model as representation of the paper: in this way we are able to access the contribution of enrichment, that is the biggest novelty of our approach. We compute the recall and the amount of human workload needed in all the simulations (i.e. the number of papers to be manually examined). We found that with a recall of 1, 83 papers are manually examined in our process, against 99 of the process without enrichment and 106 of the original manual selection. Therefore, we save, without losing any relevant paper, more than 20% of the manual work with respect to the original manual workload needed, and about 15% in the case of classification with no enrichment. In Figure 3 we represent the performance of our process in the case study (with and without enrichment), comparing it with an ideal process that selects only relevant papers. Recall is on y-axes, whilst human workload needed is on x-axes.

VI. Discussion and Threats to Validity

Two relevant construct threats are located at the begin of our process, i.e. the composition of I_0. The first one is to build an I_0 which is representative of just a niche of the field of interest. As a consequence the automated classification

[5]http://viewer.opencalais.com/

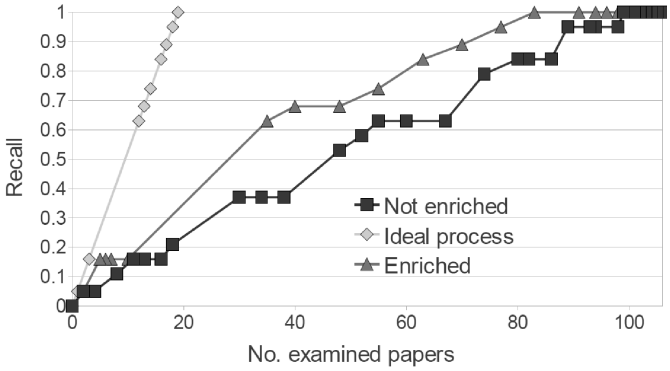

Fig. 3. The performance of the classifier with enrichment compared to the ideal process and the classifier without enrichment.

could potentially discard all resources not part of the niche described. Moreover, the second threat is the subjective bias in the composition of I_0. However, the model built on I_0 is compared to the enriched candidate source coming from W: the enrichment could lesser both the problems, and permit to correctly start the enlargement of I evading from the niche.

Proceeding on the subsequent steps of the algorithm, we identify two further threats, a construct threat in the Linked Data enrichment step and a conclusion threat in the Naive Bayes classification. The first one is the possibility that some terms are not present in DBPedia, hence a paper could not be enrichable and our approach can not be applied to it. Moreover, we could also encounter problems of terms ambiguity and synonymy. The conclusion threat is on the classification: the condition classification $P[w+_i \in I] \geq P[w+_i \in E]$ could be more precise than the condition with threshold that we select. However, we decide, for the sake of simplicity, to do not model E. Finally, the validation case study conducted is limited just to a portion of the real search of Jorgensen and Shepperd [2007], hence generalization is weak.

Despite the identified possible drawbacks, an important positive consequence of the use of automatic classification is the possibility to operate on larger search spaces because the effort of exploring W is reduced automatizing the first selection. As consequence the search strategies can also explore just remotely potential interesting sources. For example, using the standard approach, search on a high number of journals and conferences is commonly quite expansive, instead resorting on partially automatic classification this search is affordable without incurring in an overwhelming workload increment and removing the subjectivity in the classification.

Moreover, using Linked Data we are able to capture not just papers we recognize being similar to the ones already selected but we are able to capture papers that have conceptual relations to the content expressed in the already selected papers. This strategy permits to deal with an incomplete description of the field of interest, which can not be completely described by the set of already selected papers. Therefore Linked data permits to use our approach also with an I set which is relative small and not representative of the whole field. Finally we are confident that results from our case study are sensible because the dataset used was quite dense (24 papers relevant out of 111). This is more dense than the typical search space normally used in systematic reviews, so repeating the same case study on more sparse datasets it is reasonable to yield a greater workload reduction.

VII. Conclusion and Future Work

In this paper we present an improvement to the standard approach for performing selection of studies in a systematic review. We explain how our approach differs from the one presented by Kitchenham [2004] and we list two main advantages of an enriched selection process: i) a reduction of workload requested to classify sources and ii) a reduction of subjectivity in the overall process. We conducted a case study to compare our process with a traditional approach: we obtained a good reduction of work load, without losing any relevant paper. As future work we desire to deal with threats to validity and to conduct a wider empirical validation of this process.

References

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. ISSN 1570-8268.

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. ISSN 1067-5027.

M. Jorgensen and M. Shepperd. A systematic review of software development cost estimation studies. *Software Engineering, IEEE Transactions on*, 33(1):33–53, January 2007. ISSN 0098-5589.

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

F. Peinemann, N. Mcgauran, S. Sauerland, and S. Lange. Disagreement in primary study selection between systematic reviews on negative pressure wound therapy. *BMC Medical Research Methodology*, 8:8–41, June 2008. ISSN 1471-2288.

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.

K. Schneider. A comparison of event models for naive bayes anti-spam e-mail filtering. In *Proc. of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 307–314. Association for Computational Linguistics, 2003. ISBN 1-333-56789-0.

M. Staples and M. Niazi. Experiences using systematic review guidelines. *J. Syst. Softw.*, 80:1425–1437, September 2007. ISSN 0164-1212.