Mining Semantic Relations between Research Areas

Francesco Osborne[1], Enrico Motta[2]

[1]Dept. of Computer Science, University of Torino, 10149 Torino, Italy
osborne@di.unito.it
[2]Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
e.motta@open.ac.uk

Abstract. For a number of years now we have seen the emergence of repositories of research data specified using OWL/RDF as representation languages, and conceptualized according to a variety of ontologies. This class of solutions promises both to facilitate the integration of research data with other relevant sources of information and also to support more intelligent forms of querying and exploration. However, an issue which has only been partially addressed is that of generating and characterizing semantically the relations that exist between research areas. This problem has been traditionally addressed by manually creating taxonomies, such as the ACM classification of research topics. However, this manual approach is inadequate for a number of reasons: these taxonomies are very coarse-grained and they do not cater for the fine-grained research topics, which define the level at which typically researchers (and even more so, PhD students) operate. Moreover, they evolve slowly, and therefore they tend not to cover the most recent research trends. In addition, as we move towards a semantic characterization of these relations, there is arguably a need for a more sophisticated characterization than a homogeneous taxonomy, to reflect the different ways in which research areas can be related. In this paper we propose Klink, a new approach to i) automatically generating relations between research areas and ii) populating a bibliographic ontology, which combines both machine learning methods and external knowledge, which is drawn from a number of resources, including Google Scholar and Wikipedia. We have tested a number of alternative algorithms and our evaluation shows that a method relying on both external knowledge and the ability to detect temporal relations between research areas performs best with respect to a manually constructed standard.

Keywords: Research Data, Ontology Population, Bibliographic Data, Empirical Evaluation, Scholarly Ontologies, Data Mining.

1 Introduction

Consistently with the general trend towards characterizing information using Semantic Web standards, for a number of years now we have seen the emergence of repositories of research outputs specified using OWL/RDF as representation languages – e.g., see [1], [2], [3], and conceptualized according to a variety of ontologies, such as SWRC[1], BIBO[2], and AKT[3]. This class of solutions promises both to facilitate the integration of research data with other relevant sources of information – e.g., data drawn from social media [4], and also to support more intelligent forms of querying and exploration. In particular, in order to make sense of the key trends and dynamics of a research area, it is essential to have tools which are able to support a seamless exploration of the various relations that exist between authors, publications, impact measures, publication venues, research areas, etc. Within this context, it is particularly important to associate correctly authors and publications to research areas, to ensure good precision and recall when exploring what goes on within a particular research area.

The association between authors and publications on the one hand and research areas on the other is normally determined on the basis of the keywords that authors themselves associate with their publications. However, this purely syntactic approach is unsatisfactory for a number of reasons: authors do not necessarily use a consistent terminology to specify the relevant research areas and, even when they do, a syntactic approach fails to capture the relations that may exist between research areas – e.g., most researchers consider "ontology alignment" and "ontology matching" as essentially equivalent labels for the same research area, but searching for "ontology alignment" in most bibliographic servers does not return papers tagged as "ontology matching". Hence, there is a need for methods which are able to generate the relations which exist between research areas, to enable more intelligent querying and exploration of research data.

This problem has been traditionally addressed by manually creating taxonomies, such as the ACM classification[4]. However, this manual approach suffers from a number of problems. These taxonomies are very coarse-grained and they do not cater for the fine-grained research topics, which define the level at which typically researchers (and even more so, PhD students) operate. Moreover, because these taxonomies are defined manually, they evolve slowly, and therefore they do not cover the most recent research trends. In addition, as we move towards semantically characterized repositories of research data, there is arguably a need for a more sophisticated representation of the relations between research areas, than a homogeneous and un-typed taxonomy, to reflect the different ways in which research areas can be related.

In this paper we address this problem by proposing Klink, a new approach to automatically generating relations between research areas, which combines both machine learning methods and external knowledge, drawn from a number of resources, including Google Scholar and Wikipedia. In particular, we have tested a number of alternative algorithms and our evaluation shows that a method relying on both external knowledge and temporal relations between research areas performs best with respect to a manually constructed standard and indeed achieves a very good level of precision and recall.

[1] http://ontoware.org/swrc/.
[2] http://bibliontology.com.
[3] http://www.aktors.org/publications/ontology.
[4] http://www.acm.org/about/class/ccs98-html.

2 What's in a link: characterizing relations between research areas

Taxonomies of research areas are not like taxonomies in other domains, in the sense that there is not necessarily an all-encompassing and 'objective' organization of research topics. For example, one of the authors of this paper was involved in one of the very first attempts at building a semantic repository of research data, the KA2 initiative [5], and participated in a workshop whose main goal was to generate a taxonomy of research topics. This turned out to be much harder than predicted, given that for a number of topics there were serious disagreements about their relationships with other topics. Nevertheless, it is also the case that, given a research community, there are typical many relatively unproblematic cases where a broad consensus can be found about an area being equivalent to or being a sub-area of another area. For instance, we earlier made the example of the terms "ontology alignment" and "ontology matching" being used practically as synonyms in the research community. Another relatively uncontroversial example concerns the area of Semantic Web Services, which most people agree is a sub-area of both Web Services and Semantic Web.

However there are also other situations that are rather less obvious. For instance, while there may be a certain degree of consensus that research in Ontology Engineering is relevant to the Semantic Web area, most people would disagree with the statement that Ontology Engineering is a sub-area of Semantic Web. Nevertheless, if I am looking for papers on the Semantic Web, it may actually be useful for me if my system for research data exploration were also able to flag papers in Ontology Engineering as potentially relevant. And indeed, the relevance of the latter area of research to the former can be easily ascertained by browsing the proceedings of the main Semantic Web conferences.

In sum, the point here is that simply looking either for strict equivalence between research areas or strict *subAreaOf* relations is unsatisfactory, because it may fail to capture some other useful relations between research areas. For this reason, in our work so far we have also included relations such as that exemplified by Ontology Engineering and Semantic Web, where the results from the former contribute to research in the latter. Hence, our model at the moment considers the following three relations between research areas:

- *relatedEquivalent*. This is defined as a sub-property of *skos:related*, which indicates that two particular ways of referring to research areas can be treated as equivalent for the purpose of exploring research data – e.g., "ontology alignment" and "ontology matching" can be considered as equivalent.
- *skos:broaderGeneric*. We reuse this property from the SKOS[5] model, to indicate that a research area – e.g., Web Services, is broader than Semantic Web Services. Transitivity is important here, because this property is used to characterize the intuitive notion that an area is a sub-area of another one.
- *contributesTo*. This is defined as a sub-property of *skos:related* and indicates that while an area, R1, is not a sub-area of another one, R2, its research outputs

[5] http://www.w3.org/2004/02/skos/.

are so relevant to R2 that it may be useful for the purposes of querying and exploration to assert this relationship, to provide better support to users.

However, it is important to emphasize that, while our epistemology distinguishes between the aforementioned three relations, the current version of our algorithm, which will be presented in the next section, is only able to differentiate automatically between *hierarchical* and equivalent relations. In other words, while the algorithm is able to differentiate *relatedEquivalent* relations from the others, and it is also able to mine both *contributesTo* and *skos:broaderGeneric* relations, it treats these two relations as generic hierarchical relations and cannot differentiate them further. Hence, this final step – i.e., separating *contributesTo* from *skos:broaderGeneric* relations, needs at the moment to be carried out manually.

Our model[6] builds on the BIBO ontology, which in turn builds on SKOS[7], FOAF[8], and other standards. Our goal here was not to produce yet another ontology, so our extensions to BIBO are very conservative and comprise only the *relatedEquivalent* and *contributesTo* object properties described earlier, and the class *Topic*, which is used to refer to research topics.

3 The Klink algorithm: automatically detecting relations between research areas

3.1 Preliminaries

We propose a novel approach, named Klink, for cleaning and inferring hierarchical and equivalence relationships from a set of keywords associated with a collection of documents.

Klink detect links between keywords by using heuristic rules, statistical methods and external knowledge. Moreover it allows a human user to define some aspects of the hierarchy, such as the maximum permitted number of parent nodes for each node. An important aspect of Klink is that it is able to discard keywords which are not research areas but can be used as keywords for a paper. Typical examples include names of software tools as well as ‘orthogonal’ keywords, e.g., “Case Study”, which do not denote a research area but a particular aspect of the paper in question – i.e., that a case study is presented.

[6] http://kmi.open.ac.uk/technologies/rexplore/ontologies/BiboExtension.owl.

[7] The most recent specification of the SKOS model, which can be found at http://www.w3.org/TR/2009/REC-skos-reference-20090818/, makes a number of modifications to the modeling of these relations and in particular proposes a new property, *skos:broaderTransitive,* to support the representation of transitive hierarchical relations. Here we stick to the older SKOS specification, primarily because our conceptual model builds on the BIBO ontology, which in turn builds on the 2004 SKOS model. While there are interesting semantic differences between the different versions of the SKOS model, in the context of this paper these are not so important, as we are only concerned with extracting the three kinds of relations between research areas, which have been presented above.

[8] http://xmlns.com/foaf/spec/.

Since we use a statistical approach it is imperative to have an unbiased and large enough collection of documents. To do any kind of inference on a keyword that has a low number of occurrences may be risky; it is better to discard it, at the cost of losing some useful piece of information. We should also be careful not to introduce biases when extracting subsets from a larger population. For example if we were to analyze a sample composed only by papers from the five best Semantic Web conferences, the importance of Semantic Web with respect to other areas would be necessarily overestimated. In that sample we may in fact discover that 80% of the papers about Machine Learning are associated with Semantic Web, and thus erroneously conclude that Machine Learning is a sub-area of Semantic Web. For this reason, while our experiments zoom on the Semantic Web as the 'focus topic', the corpus we use is very large and includes more than one million and half papers downloaded from Microsoft Academic Search[9] (MAS), which by and large are situated in the Computer Science area.

3.2 Overview of the approach

The input to Klink is a collection of keywords associated with a set of documents and the result is a graph structure containing both hierarchical and equivalence links. The outline of the algorithm is as follows:

1) Each keyword in input is compared to the other keywords with which it shares at least *n* co-occurrences and two kinds of hierarchical links are inferred: the *'standard'* one and the *'temporal'* one;
2) Each keyword is checked for possible deletion if it does not meet the requirements for being a research area;
3) The links are cleaned by deleting triangular and circular hierarchical relationships and the eventual user's requirements on the structure are enforced;
4) Each keyword is compared to the other keywords with which it shares at least *n* co-occurrences; the *relatedEquivalent* relationships are inferred and the relative keywords are merged;
5) Step 1, 3 and 4 are repeated with the new keywords obtained by merging the keywords with inferred equivalence relationships until no new *relatedEquivalent* relationships emerge.

It should be noticed that step 2 will be run only once and, as a choice, can be applied after step 5, giving the keywords that should be deleted the possibility of entering into a *relatedEquivalent* relationship.

3.3 Step 1 – Inferring hierarchical relationships

In the classical definition of subsumption [6], term *x* is said to subsume term *y* if two conditions hold: $P(x|y) = 1$ and $P(y|x) < 1$, e.g. if *y* is associated to documents that are a subset of the documents *x* is associated to. Usually the first condition is relaxed in $P(x|y) > \alpha$, since it is quite improbable to find a perfect relationship. The usual value of α is 0.8, although other values are possible according to the kind of documents

[9] http://academic.research.microsoft.com.

examined. For the inference of a hierarchical relationship between keywords we use a variation of this idea, combined with other heuristic metrics. We consider two different kinds of links, the standard one and the temporal one.

3.3.1 Inferring standard hierarchical links

We define as a hierarchical link of x with respect to y the relationship in which the difference between $P(y|x)$ and $P(x|y)$ leans decidedly toward y and the two terms co-occur with a similar set of keywords.

We compute the strength of the hierarchical relationship as:

$$L(x,y) = (P(y|x) - P(x|y)) * c(x,y) * (1+ N(x,y))$$

where $c(x, y)$ is the cosine similarity between keywords and $N(x,y)$ is a metric that weighs the similarity of the keyword names. This similarity is computed as the ratio between the number of identical words between two keywords and their average number of words.

A hierarchical link is inferred when $L(x,y) > t$, and thus x is considered a sub-area of y. We suggest a value of 0.2 for t, and in the evaluation we will show how recall and precision change for different values of t. It is also possible to use other additional filters, chosen carefully according to the set of documents. We actually experimented with some of them on the sample of metadata downloaded from MAS, obtaining interesting results. Specifically we used the condition that a keyword had to be at least two years older than another as a necessary (but not sufficient) condition for being considered as its super-area. We then experimented with a filter based on dimensions, accepting as sub-areas only areas n times smaller then the super-area. However this technique can bring more problems than advantages, since an area can outgrow its super-area. Our conclusion was that this filtering technique might be useful but it is strongly dependent on the characteristics of the selected collection of documents.

3.3.2 Inferring temporal hierarchical links

Some relationships among areas may escape the mechanism previously described. Usually, when an area is mature enough, references to its super-area become implicit and no longer appear as co-occurring keywords. For example many disciplines fall under the Artificial Intelligence area, but today it is hard to find explicit references to it in the keywords associated with a publication. For a human it is not very informative to annotate that Machine Learning is a form of Artificial Intelligence, since it is already a huge and independent research area by itself. The information about the origins of a research area is however necessary when building a complete taxonomy.

As an area grows, the co-occurrences with its super-area become fewer and fewer, making harder to infer the origins of a topic by looking only at the total co-occurrences. Taking into consideration also the temporal dimension aims to solve this drawback. The idea is that the initial co-occurrences of two keywords are the most informative about stating that a parent area somehow spawned a sub-area.

We use the term *temporal link* to refer to the relationship behind this intuition. It should be noted that, although temporal links can be used together with the standard links to build a taxonomy, they have a different meaning. A standard link between x and y implies that y was a vital keyword for x along the total life of x. The temporal link instead implies that the set of keywords with which x co-occurred in its first years are privileged. Therefore a temporal link between x and y means that y was a very important keyword for x in the initial years of x's life. As time goes by, the relationship with the topic inferred by means of the temporal link may persist, or become implicit or even vanish.

The temporal weighted co-occurrence $CO_t(x,y)$ is obtained by adding over the years the number of co-occurrences weighted by a factor $w(year,x)$ given by:

$$w(year,x)= (debut(x) - year)^{-\gamma}$$

where $debut(x)$ is the year of the debut of keyword x, $year$ is the year in which a co-occurrence occurs, and γ is a constant > 0 that modules the importance of co-occurring in a given year. We empirically set this value to 2. It is advisable not to use as debut the first year in which a keyword appeared, but rather the first year in which it appeared in at least a minimum number of papers. We use as limit 30 papers, but according to our tests any number between 10 and 50 gives reasonable results. To reduce the noise it is also possible to take in consideration only the first n years.

The temporal subsumption metrics $L_t(x,y)$ is computed as for the standard link, using the temporal conditional probability $P_t(x|y)= CO_t(y,x)/ CO_t(y,y)$:

$$L_t(x,y) = (P_t(y|x) - P_t(x|y)) * c(x,y) * (1+N(x,y))$$

As before, a temporal link is inferred if $L_t(x,y)>t_t$. As for the standard link, we suggest a threshold value of 0.2, and in the evaluation we will show the outcome for different values of t_t.

3.3.3 Integrating external knowledge

Using external knowledge to integrate the examined corpus of documents is not always necessary but it can be very useful.

We often want to build a general taxonomy that reflects more the common use of a keyword in a certain domain rather than its interpretation in a particular set of documents. The examined set may in fact use keywords with a non-common meaning or be biased in some way. In this case it is advisable to rely on a neutral source of information [7]. Moreover, as it will be shown in section 3.4, external knowledge is vital to discover and discard keywords that do not belong to a certain domain.

We focused on the knowledge of the dimension of a keyword and of the co-occurrences of keyword pairs in different online sources. By choosing the right sources we can be sure to obtain this knowledge within a given domain – e.g., the academic one. We used parsers to collect this kind of knowledge from Google[10], Google Scholar[11], Wikipedia[12], and Eventseer[13].

Google and Wikipedia give a good approximation of keyword presence in general. On the contrary, Google Scholar focuses on the academic domain and in particular on the title and abstract of papers. Eventseer is a site that collects calls for papers, and as a result is very useful for understanding the dynamics among keywords as conferences topics. We used Google to search in both Wikipedia and Eventseer.net since the internal search of these services do not support the AND conjunction. We

[10] www.google.com

[11] scholar.google.com

[12] www.wikipedia.org

[13] eventseer.net

then exploited the "About [...] results" text for estimating the occurrence of a keyword. For the co-occurrence we used the same technique, using the AND between them in the query. We also experimented with the OR conjunction, but the combination of AND with OR seemed to yield inconsistent results.

We computed the external probability as the weighted average of the probability of a co-occurrence for each different source: $P_{ex}(x|y) = \Sigma_i w_i P_i(x|y)$. In the evaluation we considered Wikipedia (w=0.2), GoogleScholar (w=0.4) and Eventseer (w=0.4) and we computed the hybrid probability as

$P_h(x/y)= w_h P(x/y) + (1-w_h)P_{ex}(x/y)$

where $0<w_h<1$ is the constant that reflects the importance of the external probability. We set this value to 0.5 to balance the contributions of the two components. When the set of document is not very large it may instead make sense to rely more heavily on external knowledge. We can then compute the hybrid version of $L(x,y)$ by simply using the hybrid probability $P_h(x|y)$ instead of the standard one.

It is important to mention that it is not possible to use external knowledge from the aforementioned sources to deduce temporal links, given that these sources do not provide the distribution of the co-occurrences over the years.

3.4 Step 2 – Cleaning the keywords

When creating a taxonomy it is important to identify the keywords that are part of a certain domain, in this case the domain of 'research areas', and those that are not.

Text mining techniques may lead to noisy keywords that do not add any information and actually risk spoiling the inference process. For example in the MAS dataset we can find different keywords that are related to the academic world but is difficult to consider as research areas – e.g., "Web Pages", "Case Study", "Java Applet", and others. Hence it is important to detect and discard them from the final taxonomy. Our approach implements three techniques to filter this kind of irrelevant keywords.

The first and simplest procedure is the elimination of any keyword without inferred relationships with other keywords.

The second technique uses the distribution of the keyword co-occurrences. An acceptable keyword should have a limited set of main keywords with which it has a relatively high number of co-occurrences and then a long tail of less important one. Some keywords show instead a flatter distribution of co-occurrences over a large range of keywords. This is the case of many general words used in the academic world, such as "Case Study", which can occur in many papers on completely different topics. We identify these spurious keywords by fixing the number of main keywords and the minimum percentage of co-occurrences they should cover. If the main keyword covers too small a part of the total co-occurrences, then the keyword is discarded as being too general. In the evaluation we used as thresholds 20 and 15%, respectively.

The third technique uses external knowledge and it is basically a check on the estimated dimension of a keyword in a certain domain. To do so we compute the weighted sum of the ratios between the dimension of a given keyword and the average dimension of the keywords in the various sources:

$D_{ex}(x)= \Sigma_i w_i (D_i(x) / A_i)$

where A_i and w_i are respectively the average dimension and the relative importance of the keyword in the i-th dataset of the specific source.

Google and Wikipedia are less useful sources to consider when we want to know the dimension of a keyword in the academic world. Hence, we give more importance to conference calls (Eventseer) than to the occurrences in the title or in the abstract of a paper (GoogleScholar), by setting w_{ev}=0.6 and w_{gs}=0.4. If $D_{ex}(x)$ is below a given threshold, which we set empirically at 0.2, the keyword is dropped.

There may be keywords that have a small dimension but are nevertheless real research areas. Thus before deleting a keyword we run a check on its links: if either a normal or temporal link has a strength that is at least the double of the correspondent threshold, then the keyword is kept.

3.5 Step 3 – Cleaning the links

After step 2 we have a large number of cases in which two super-areas of a keyword are also in a hierarchical relationship. For example Word Wide Web and Semantic Web may both be super nodes of OWL whereas Word Wide Web may be also a super node of Semantic Web. Since such a taxonomy might be confusing the redundant links like the one between Word Wide Web and OWL are deleted.

In this stage, it is possible to cut away the links with lower *L(x,y,)* or $L_t(x,y)$ to satisfy the user's requirements on the maximum number of super and sub nodes.

3.6 Step 4 and 5 – Detection of *relatedEquivalent* relationships and merging of the keywords

The search for *relatedEquivalent* relationships between keywords offers many advantages. For example we can learn that "P2P" and "Peer to Peer" are actually the same topic and thus a query for any of the two will return a set of documents associated with both these keywords. Any statistical inference on the Peer to Peer area can then use a larger number of papers, and thus be more valid. *relatedEquivalent* relationships can be very important also when focusing only on building a taxonomy since they simplify the structure, making the subsumption inference easier.

A standard metric like the cosine similarity may work well in some cases but it raises two problems. The first is due to the fact that the eventual subsumption relationship between the keywords is not considered. If one of the keyword subsumes in some sense the other, a hierarchical link is preferable to a *relatedEquivalent* relationship. The second problem is that it is important to take in account the reasons why two keywords have a high cosine distance. In a taxonomy it is normal for sibling elements in the lower levels to have a high cosine similarity since they are different declinations of the same theme. Thus we need to take in consideration also the cosine similarity of the common super-areas of these keywords, namely the keywords that subsume both of them. If the cosine similarities of the two keywords with the common super-areas are comparable with their reciprocal similarity, then probably they are siblings, and are similar because they derive from the same area or areas. On the contrary, if their reciprocal similarity is higher than the one with the predecessors a *relatedEquivalent* relationship is more probable.

The metric *S(x,y)* we propose as a measure of the similarity between two keywords in a corpus of document is designed to reward the non-trivial similarities that cannot be derived from the taxonomy:

$S(x,y)= c(x,y) - w_{sa}c_{sa}(x,y) - w_{sub}|P(x|y) - P(y|x)|$

where *c(x,y)* is the cosine similarity between x and y, $c_{sa}(x,y)$ is the average cosine similarity with the common super-areas. $0<w_{sa}< 1$ weighs the effects of the common super-areas on the similarity and in the evaluation will be set at 0.2; $0<w_{sub}<1$ weighs the importance of not having a subsumption relationship. In the evaluation w_{sub}=0.2.

The last part of the formula reduces the risk of inferring a *relatedEquivalent* relationship when there is actually a hierarchical one, by introducing a *malus* correlated with the difference of the subsumption probabilities $|P(x|y) - P(y|x)|$.

We infer a *relatedEquivalent* candidate when $S(x,y)>t_{re}$, where t_{re} is the threshold chosen by a human user. In the evaluation we used t_{re}= 0.75.

In those rare occasions in which for two keywords both a *relatedEquivalent* and a standard or temporal link can be inferred, it is up to the user to decide the priority. We decided to prefer the inference of the standard link rather than the *relatedEquivalent* one, and the *relatedEquivalent* rather than the temporal link.

The keywords which are found as suitable *relatedEquivalent* candidates are processed by a bottom-up single-linkage hierarchical clustering algorithm which uses the inverse of *S(x,y)* as the distance between the elements.

The keywords in any resulting cluster are finally merged together in an aggregated keyword whose set of documents is the union of the sets of documents of all merged keywords. Since this new keyword should be inserted in the taxonomy, the process goes back to step one to start over again. The taxonomy will be considered complete and will be returned when no new *relatedEquivalent* relationships are inferred.

4 Evaluation

We used Klink to analyze a very large corpus of papers about the Semantic Web and related research areas. We needed a very big dataset that would offer challenges such as the presence of synonymous keywords to be merged after detecting a *relatedEquivalent* relationship among them and fuzzy keywords that might not be research areas. The collection of metadata available on MAS meets these requirements: moreover MAS offers useful APIs to provide access to their data.

As stated before, a rigorous analysis of a research area requires an unbiased sample of papers. Thus it would be inappropriate to take in consideration only the papers associated with the keyword “Semantic Web” or published in Semantic Web conferences. For this reason we constructed our corpus as follows: we first downloaded from MAS the metadata of 11,998 papers associated with the keyword “Semantic Web”; we then used this set to find the 120 research areas with which the “Semantic Web” has the highest number of co-occurrences and downloaded all the associated papers. The end results were 1,510,871 papers that we can consider to constitute an unbiased sample.

We tested different approaches to build a taxomomy, studying the impact of the different techniques presented in this paper on the final results. In particular we compared[14]:

1) The classic subsumption method [6] described in section 3.3 (labelled **S**);
2) The Klink approach to finding hierarchical standard links explained in section 3.3.1 (labelled **L**);
3) The Klink approach to finding hierarchical standard links with the integration of external knowledge described in section 3.3.3 (labelled **L+EXT**);
4) The full Klink algorithm, using both standard and temporal links (see section 3.3.2) with the integration of external knowledge (labelled **L+EXT+TL**).

The hypothesis was that Klink could be used to build taxonomies that are very similar, although not necessarily identical, to the ones created by a human user. Consistently with the discussion in section 2, the relationships inferred by our approach are instances of three kinds of semantic relationships: *broaderGeneric*, *contributesTo* and *relatedEquivalent*. However, as stated before, Klink is not able to distinguish between the first two relationships, characterizing both of them as hierarchical links. Hence, right now it is up to a human user to distinguish between these two types of hierarchical links, although we plan in future work to examine automatic ways to do so.

To evaluate the automatically built taxonomies we created a gold standard[15], which was passed on to three external experts for validation/revision. We started with a collection of the 120 keywords with the most numerous co-occurrences within the Semantic Web according to the MAS data. We then removed the less developed parts of the structure, e.g., the structure associated with the keyword "User Model", obtaining a final sample of 58 keywords: a reasonable size to be handled by the experts in their manual evaluation. About 15% of the relationships had to be changed to follow the directives of the experts. In about 7% of the cases the three experts disagreed on a relationship and we used the one suggested by two out of three. We chose to use a gold standard since it allowed to test not only the final version of the algorithm but also to study the different contributions offered by its parts as a function of the thresholds.

We selected two taxonomies with different degrees of focus on the Semantic Web. The first one (labeled Set1) covers “Semantic Web” together with “Formal Ontology” and “Knowledge Representation”. The second one (labeled Set 2) includes also the other areas and is expected to yield inferior results since the sample does not cover entirely those areas.

We ran the algorithms and compared the generated taxonomy with the gold standard by computing the recall and the precision of the inferred relationships and their harmonic mean (F-measure). To reduce complexity, we set the standard and the temporal link threshold at the same value.

Figure 1 shows the relation between precision and recall obtained with the four algorithms **S**, **L**, (**L+EXT**) and (**L+EXT+LT**) for the two sets.

[14] Because of space limitations, the only two parameters that we will analyze in this evaluation are the standard and temporal thresholds.

[15] The gold standard and the data generated by the algorithm are available at http://kmi.open.ac.uk/technologies/rexplore/data/.

Using the proposed metric for inferring hierarchical relationship described in section 3.3.1, **L** yields a recall of 53% for Set1 and 38% for Set2 for a precision higher than 80%. The basic subsumption method **S** found in the literature yields for the same level of precision a recall of 30% in Set1 and 8% in the Set2.

The integration of the external knowledge (**L+EXT**) appears to be effective allowing, with t=0.15, a precision of 93% and a recall of 90% for Set1 and precision 64% with recall 84% for Set2. With t=0.2, precision and recall go respectively to 95% and 69% in Set1 (78% and 64% for Set2).

The temporal links are able to improve the results even more, especially for Set2, where more difficulties are posed by the chore of inferring subtrees in areas for which we do not have the complete structure. With threshold 0.2, **(L+EXT+LT)** boosts the recall to 92% with a precision 94% for Set1 (86% and 78% for Set2). By raising the threshold to 0.25, precision reaches 98% with recall 73% (88% and 73% for Set2). Figure 2 shows the F-measure for the two sets as a function of the threshold.

All results agree in indicating **L+EXT+TL** as the best approach, followed by **L+EXT** and farther away by **L**. To statistically evaluate the differences among the curves in Figures 1 and 2, we employed the chi-square test. The comparison of precision between **L+EXT+TL** and **L** yields p=2x10^{-3} for Set1 and p=2x10^{-5} for Set2; both statistically significant. The comparison between **L+EXT+TL** and **L+EXT** yields no statistically significant differences for Set1, whereas p=9x10^{-3} for Set2. Similar results are obtained for recall. The fact that a statistically significant difference between **L+EXT+ST** and **L+EXT** exists only for Set2 indicates that the insertion of the temporal link was determinant for inferring relationships in a context where the set of keywords related to some research area is not complete.

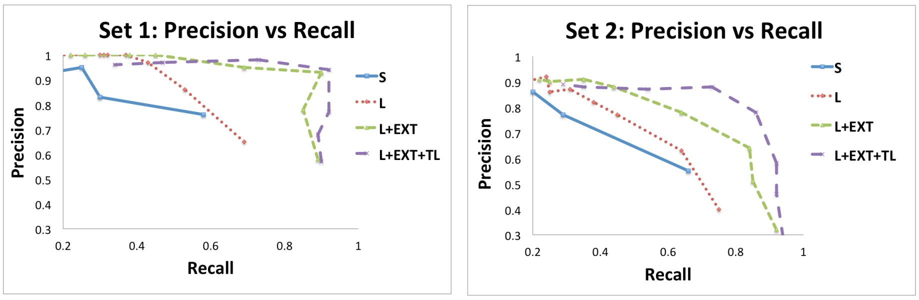

Figure 1. Precision vs Recall for Set 1 and Set 2.

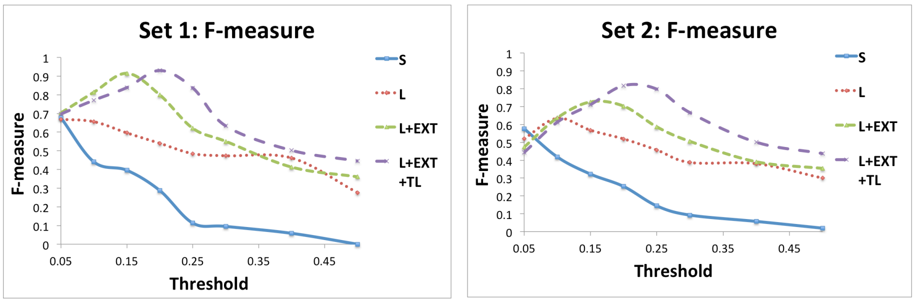

Figure 2. F-measure for Set 1 and Set 2.

The indication on the threshold places *t* between 0.15 and 0.25, depending on the desired trade off between precision and recall.

Figure 3 and Figure 4 compare the fraction of the gold standard representing the Semantic Web with the automatically generated version using the **L+EXT+LT** version of the algorithm and t=t_t=0.2.

The few discrepancies between the two figures open interesting perspectives. As an example, let us consider the relationship between the areas of Web of Data and of Linked Data. Before the test runs, two of our experts considered correct to have Web of Data as *skos:broaderGeneric* than Linked Data, whereas the third preferred *relatedEquivalent*. Our algorithm appears to have chosen a third possibility, i.e., that Web of Data is more specific than Linked Data. In our dataset the papers associated with Web of Data were also associated with Linked Data in 54% of the cases, while the contrary happens only in 7.5% of the cases. The very high value of the cosine similarity, 0.94, between the two keywords indicated a strong relationship and the co-occurrence analysis suggested that the super-area should be Linked Data. Hence, the data appear to indicate that in practice authors use "Linked Data" as a generic term for talking about the web of data to a much greater extent than they use the term "Web of Data".

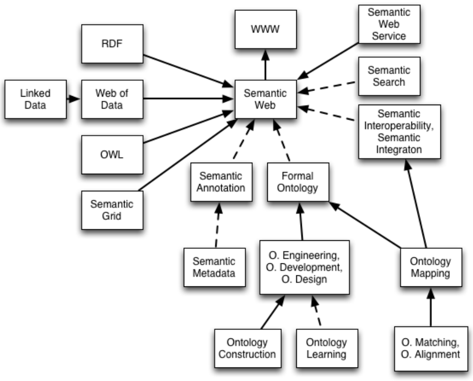

Figure 3. A portion of the gold standard associated with the Semantic Web. The solid arrows represent *broaderGeneric* relationships, the dashed ones *contributesTo* ones.

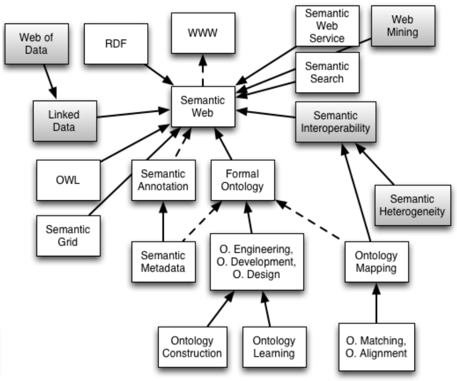

Figure 4. A portion of the automatically generated taxonomy associated with the Semantic Web. The solid arrows represent standard links, the dashed ones temporal ones.

While experts may consider this perspective incorrect, a data-driven, bottom-up approach like ours can also be used to highlight these interesting discrepancies between the intended coverage of research areas and the mental models that emerge from the way these terms are used in.

5 Related Work

Taxonomies can be very useful tools for improving search results and their presentation [8]. In particular these structures are helpful in faceted search [9], which is the search paradigm based on the indexing of documents along multiple orthogonal taxonomies. Faceted semantic search systems tend instead to use ontological relationships, such as *partOf* or *isA* [10].

To build manually a taxonomy is however not a trivial task and the human crafted ones may not mirror the true internal relationships of a corpus of documents, but rather the standard taxonomy of the field. For this reason to automatically generate a taxonomy from a corpus of document or metadata is an important challenge.

Traditionally there have been two different approaches to this task. The first is based on clustering techniques [11], the second, developed especially in computational linguistic, rests on the detection of lexico-syntactic patters [12].

The *TaxGen* framework [13] uses a hierarchical agglomerative clustering algorithm and text mining techniques for the creation of a taxonomy from a collection of unstructured documents. In [14] a hierarchical clusterization algorithm is applied on web pages and a top-down partitioning is used to generate a multi-way-tree taxonomy from the binary tree. Our method also exploits hierarchical algorithms and similarity distances between keywords. However they are not used for generating the hierarchical structure but for merging the keywords for which a *relatedEquivalent* relationship was inferred.

The other traditionally used method is based on the detection of linguistic patterns that appear in a corpus of documents. In [15] an approach called Lexico-Syntactic Pattern Extraction (LSPE) is presented, which exploits patterns like "such as…" and "and other…" to discover relationships between terms. A similar procedure is also reported in [16] where a clustering-based sense disambiguation heuristics is proposed for pruning the resulting taxonomy. The same technique can be used also to infer ontological relationships like *subClassOf*, as in [17].

The work of Sanderson and Croft [6] proposes an approach that allows generating automatically concept hierarchies without the use of training data or clustering techniques. Namely they use the probability that a keyword is associated with another to infer subsumptions, as discussed in section 3.3. The same idea is extended in the GrowBag algorithm [18], which exploits the second order co-occurrence made explicit by a biased *PageRank* algorithm.

The basic idea that we have used to infer the subsumption relationship between keywords is similar to the one found in [6]. However we extended this approach i) by introducing a set of very different metrics, which exploit cosine similarities and temporal dimensions and ii) by integrating external knowledge into the process.

Other authors have proposed the use of external knowledge from web pages for finding hierarchical relationships. In [7] three heuristic techniques are suggested for mining topic-specific knowledge, however such methods need specific patterns that may not be very common in all domains.

The approach proposed in this paper can find practical applications in the growing areas of academic repositories (see for example [3], [19]), to support users in the exploration and use of such repositories. In particular, it can be seen as a complementary approach to techniques employed for managing, cleaning and organizing folksonomies of tags attached to research papers, as notably applied in the Bibsonomy system [20]. The semantic GrowBag algorithm already mentioned was similarly employed to derive automatically facets to be used in the faceted browsing of large publication collections (in Faceted DBLP, see [18]). Our approach however focuses on finding relationships between keywords with a high level of accuracy, which are verified as corresponding to research areas. As a result we can provide a robust navigational structure for collections of research publications, while reducing the need for manually curating the structure of the collection.

Other related works concern complementary approaches, which investigate the connections between authors of papers (the network of researchers), in order to establish relationships in their areas of interest (see for example [21]). The results obtained naturally differ in their views of the types of relationships shared between research areas/communities.

6 Conclusions

We have presented Klink, a novel algorithm to infer relationships between keywords from a collection of terms associated with documents. Klink was tested on a large corpus of data from MAS to analyze the relationships between the Semantic Web research area and other related areas.

The results of the evaluation shows a statistically significant improvement of the performance by using Klink over the classic subsumption method: the values of recall and precision obtained in regard to the gold standard are highly satisfactory. The keyword-centered perspective of the algorithm also offers interesting opportunities for analyzing situations in which the experts do not agree on the kinds of relationships between two research areas.

The next steps include three main avenues of work. Firstly, we are currently developing a novel system, called *Rexplore*, whose aim is to improve the support available to users to explore and make sense of research data, by integrating a wide range of novel visualization methods. The method presented in this paper will be integrated with Rexplore, to support a more powerful and flexible way to map research areas to authors and publications. The second avenue of work focuses on developing new methods to automatically distinguish between *broaderGeneric* and *contributesTo* relationships, to avoid the need for humans to perform this final semantic step. Finally we want to improve our algorithm by allowing it to recognize sets of keywords of similar meaning that fall under a common area, even when this is not explicitly present in the keyword collection. For example, both RDF and OWL can be seen as sub-areas of a more generic area, which could be called "Web Knowledge Representation", however such area is rarely used as a keyword by authors. Again, by using a combination of machine learning techniques and external knowledge, we are confident that a method can be developed, which will be able to handle these situations correctly.

References

1. Moller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for Semantic Web Dog Food — The ESWC and ISWC Metadata Projects. In: 6th International Semantic Web Conference, 11-15 Nov 2007, Busan, South Korea. (2007)
2. Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H.: Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). In: WWW'2010 Workshop on Linked Data on the Web (LDOW 2010), CEUR-WS Vol-628, Raleigh, North Carolina, USA. (2010)

3. Glaser, H., Millard, I.: Knowledge-Enabled Research Support: RKBExplorer.com. Proceedings of Web Science 2009, Athens, Greece. (2009)
4. Stankovic, M., Rowe. M.: Mapping Tweets to Conference Talks: A Goldmine for Semantics. ISWC 2010 Workshop on Social Data on the Web, Shanghai, China. (2010)
5. Benjamins, R.; Fensel, D.: and Decker, S.: KA2: Building Ontologies for the Internet: A Midterm Report. International Journal of Human-Computer Studies 51(3). (1999)
6. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In Proceedings of the SIGIR conference, pp. 206–213. (1999)
7. Liu, B., Chin, C. W., Ng, H. T.: Mining topic-specific concepts and definitions on the web. Proceedings of WWW 2003, pp. 251-260. ACM, New York, USA. (2003)
8. Pratt, W., Hearst, M.A., Fagan, L.M.: A knowledge-based approach to organizing retrieved documents. AAAI conference, Menlo Park, CA, USA. (1999)
9. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A browser for heterogeneous semantic web repositories. Proceedings of the 5th Int. Semantic Web Conference, vol. 4273/2006 of LNCS, p. 272–285. Springer Berlin/Heidelberg. (2006)
10. Suominen, O, Viljanen, K., Hyvänen, E.: User-Centric Faceted Search for Semantic Portals. Proceedings of the 4th European conference on The Semantic Web: Research and Applications (ESWC '07) pp. 356-370. (2007)
11. Assadi, H.: Construction of a regional ontology from text and its use within a documentary system. In N. Guarino (ed.), Formal Ontology in Information Systems, Proceedings of FOIS-98, pp. 236-249. Trento, Italy. (1999)
12. Morin, E.: Automatic acquisition of semantic relations between terms from technical corpora. Proceedings of the 5th International Congress on Terminology and Knowledge Engineering. (1999)
13. Müller, A., Dorre, J.: The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection. Proceedings of the 32nd Hawaii International Conference on System Sciences-Volume 2, pp. 20-34. (1999)
14. Chuang, S., Chien, L.: A practical web-based approach to generating topic hierarchy for text segments. Proceedings of the 13th ACM Conference on Information and Knowledge Management. Washington, D.C., USA. (2004)
15. Hearst, M.: Automated discovery of WordNet relations. In C.Fellbaum, WordNet: An Electronic Lexical Database, pp. 131-153. MIT Press. (1998)
16. Recio-Garcia,J., Wiratunga, N.: Taxonomic semantic indexing for textual case-based reasoning. Proceedings of ICCBR 2010, pp. 302-316. Springer-Verlag. (2010)
17. De Cea, G., de Mon, I., Montiel-Ponsoda, E.: From Linguistic Patterns to Ontology Structures. 8th Conference on Terminology and Artificial Intelligence. (2009)
18. Diederich, J., Balke, W., Thaden, U.: Demonstrating the Semantic GrowBag: Automatically Creating Topic Facets for FacetedDBLP. Proceedings of JCDL '07, ACM, New York, NY, USA. (2007)
19. Jaschke, R., Grahl, M:, Hotho, A., Krause, B.:, Schmitz, C. and Stumme, G.: Organizing Publications and Bookmarks in BibSonomy. WWW Workshop on Social and Collaborative Construction of Structured Knowledge. (2007)
20. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C, Stumme, G.: The social bookmark and publication management system bibsonomy. VLDB Journal, 19(6), pp. 849-875. (2010)
21. Krafft, D., Cappadona, N., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.: VIVO: Enabling National Networking of Scientists. Proceedings of the Web Science Conference 2010, pp. 1310-1313. Raleigh, US. (2010)