Tag Recommendation for Large-Scale Ontology-Based Information Systems

Roman Prokofyev[1], Alexey Boyarsky[2 3 4], Oleg Ruchayskiy[5], Karl Aberer[2], Gianluca Demartini[1], and Philippe Cudré-Mauroux[1]

[1] eXascale Infolab, University of Fribourg—Switzerland
{firstname.lastname}@unifr.ch

[2] Ecole Polytechnique Fédérale de Lausanne—Switzerland
{firstname.lastname}@epfl.ch

[3] Instituut-Lorentz for Theoretical Physics, U. Leiden—The Netherlands

[4] Bogolyubov Institute for Theoretical Physics, Kiev—Ukraine

[5] CERN TH-Division, PH-TH, Geneva—Switzerland
oleg.ruchayskiy@cern.ch

Abstract We tackle the problem of improving the relevance of automatically selected tags in large-scale ontology-based information systems. Contrary to traditional settings where tags can be chosen arbitrarily, we focus on the problem of recommending tags (e.g., concepts) directly from a collaborative, user-driven ontology. We compare the effectiveness of a series of approaches to select the best tags ranging from traditional IR techniques such as TF/IDF weighting to novel techniques based on ontological distances and latent Dirichlet allocation. All our experiments are run against a real corpus of tags and documents extracted from the ScienceWise portal, which is connected to ArXiv.org and is currently used by growing number of researchers. The datasets for the experiments are made available online for reproducibility purposes.

1 Introduction

The nature of scientific research is drastically changing. Fewer and fewer scientific advances are carried out by small groups working in their laboratories in isolation. In today's data-driven sciences (be it biology, physics, complex systems or economics), the progress is increasingly achieved by scientists having heterogeneous expertise, working in parallel, and having a very contextualized, local view on their problems and results. We expect that this will result in a fundamental phase transition in how scientific results are obtained, represented, used, communicated and attributed. Different to the classical view of how science is performed, important discoveries will in the future not only be the result of exceptional individual efforts and talents, but alternatively an emergent property of a complex community-based socio-technical system. This has fundamental implications on how we perceive the role of technical systems and in particular information processing infrastructures for scientific work: they are no longer a subordinate instrument that facilitates daily work of highly gifted individuals, but become an essential tool and enabler for performing scientific progress, and eventually might be the instrument within which scientific discoveries are made, represented and brought to use.

Any such tool should in our opinion possess two central components. One is a *field-specific ontology*, i.e., a structured organization of the knowledge created by the researchers in a given field, along with a formal description of the information and processes they utilize. While in some important cases (e.g., in bioinformatics or chemistry) it is possible to create large ontologies of sufficiently homogeneous concepts and automatically manipulate them using formal rules (see e.g. [13]), the ontology of scientific knowledge *per se* is very complex and vaguely defined at any given point in time. Scientific ontologies can therefore only be created by a combination of existing automatic methods and novel approaches that will enable human-machine collaboration between scientists and the knowledge management infrastructures allowing to combine presentation of new results, in-depth discussions, "user-friendly" introductions for young scientists, and meta-data to relate semantically similar concepts or pieces of content. Today, there are no standard tools to insert, store and query such meta-data online, which mostly remains "in the heads of the experts" [1].

The organization of scientific information does not end with the generation of the scientific ontology. The second crucial element is a set of meaningful connections between such an ontology and the body of research material (papers, books, datasets, etc.). The challenge here is to connect semi-structured data to the natural language content of scientific papers through semantically meaningful relations. This creates a number of challenges to the current state-of-the-art in information retrieval, entity recognition and extraction (since scientific concepts can have many different names and context-dependent meanings).

In this paper, we tackle the problem of *ontology-based tagging*, i.e., of improving the relevance of automatically selected tags in large-scale ontology-based information systems. Contrary to traditional settings where tags can be chosen arbitrarily, we focus on the problem of recommending tags (e.g., concepts) directly from a collaborative, user-driven ontology.

The contributions of this paper are as follows:

- We formally define the task of ontology-based tagging and suggest standard metrics borrowed from Information Retrieval to evaluate it.
- We contribute a real document collection, a domain-specific ontology, and lists of expert-provided tags picked from the ontology and assigned to the documents as a standard evaluation collection for ontology-based tagging.
- We compare the effectiveness of standard Information Retrieval techniques (based on Term Frequency and Inverse Document Frequency) on our evaluation collection.
- We also compare the effectiveness of ontology-based techniques (e.g., based on ontological neighborhood or subsumption) and semantic clustering techniques (such as Latent Semantic Indexing and Dirichlet Allocation).
- Finally, based on the results of our experiments, we draw conclusions w.r.t. the practically and usefulness of using a given technique for ontology-based tagging and discuss future optimizations that could be used to improve our results.

The rest of this paper is structured as follows: We start by discussing related work in Section 2. We briefly present ScienceWise, the infrastructure we leverage on for our experiments, and formally define the task we tackle in Section 3. We discuss our metrics and data sets in Section 4. We report on our experimental results and compare

the effectiveness of various approaches for ontology-based tagging in Section 5, before concluding in Section 6.

2 Related Work

Research on tag recommendation can be classified into two main categories. A first class of approaches look at the contents of the resources while a second type look at the structure connecting users, resources, and tags. Examples of the former class include content-based filtering [11] and collaborative-filtering tag suggestion techniques [17]. Along similar lines, we previously experimented with tag propagation in document graphs in [6]. The latter class includes approaches that focus on the user rather than just providing tag recommendations given a resource. In [10] a set of candidate tags is created and then filtered based on choices made by the user in the past. An approach based on a user-resource-tag graph is FolkRank [8]: It computes popularity scores for resources, users, and tags based on the well-known PageRank algorithm. The assumption is that importance of resources and users propagates to tags.

Word sense disambiguation (WSD) is the task of identifying the correct meaning of an ambiguous word (e.g., 'bank' can indicate either a financial institution or a river bank). A common technique for WSD is to exploit the context of ambiguous words, that is, other words in its vicinity (e.g., in the same sentence). An approach following this idea has been used by Semeraro et al. in [4] where among all the possible senses for a word in WordNet [7], the correct one is chosen by measuring the distance (based on text similarity functions) between the word context and its synsets (i.e., the set of all synonyms for one sense).

Though tag recommendation and disambiguation have been studied extensively (both for free-text tagging and folksonomy systems), surprisingly little research has been carried-out for tag recommendation and disambiguation in a Semantic Web context. Contag [3] is an early system recommending tags by extracting topics using online Web 2.0 services and matching them to an ontology using string similarity. To the best of our knowledge, the present effort is the first systematic and repeatable experimental study of tag recommendation for large-scale and collaborative ontology-based information systems.

3 The ScienceWISE system

The ScienceWISE system allows a community of scientists, working in a specific domain, to generate dynamically as part of their daily work an *interactive semantic environment*, i.e., a field-specific ontology with direct connections to research artifacts (e.g., research papers) and scientific data management services. The central use-cases of ScienceWISE are *annotations* (e.g., adding "supplementary material" or meta-data to scientific artifacts) and *semantic bookmarking* (e.g., creating virtual collections of research papers from `ArXiv.org` [2]).

The system has been public for about one year and is accessible by scientists via our website[6], as well as via `ArXiv.org` and the CERN Document Server[7]. The system cur-

[6] `http://sciencewise.info/`
[7] `http://cds.cern.ch`

rently counts above 200 *active users* (using our services on a regular basis), thousands of annotated papers, and is now receiving several new registrations *daily*.

The domain-specific ontology is central to our system and allows us to integrate all heterogeneous pieces of data and content shared by the users. Since the underlying domain of the ontology is often rapidly changing and only loosely-defined, the best way to keep it up to date is to crowdsource its construction through the community of expert scientists. To create the initial version of the ontology, we have performed a semi-automated import from many science-oriented ontologies and online encyclopedias. After this initial step, ScienceWISE users (who are domain experts) are allowed to edit elements of the ontology (e.g., adding new definitions or new relations) in order to improve both its quality and coverage. Presently, the ScienceWISE ontology counts more than 60'000 unique entries, each with its own definitions, alternative forms, and semantic relations to other entries.

In the context of this paper, we focus on two important and related problems that we have to tackle in order to improve the user experience: tag recommendation and tag disambiguation. We note that those two tasks are key not only in our setting, but for all large-scale, collaborative and ontology-based information systems that are currently gaining momentum on the Internet.

3.1 Tag Recommendation

When users bookmark an `ArXiv.org` paper, our system attempts to automatically select the most relevant tags for characterizing the paper. The tags in question are in our case scientific concepts that are defined in the ontology. A user-friendly interface allows then to correct the system recommendation, e.g., by adding relevant tags or removing irrelevant tags from the top-k list that the system recommended.

More formally, the tag recommendation task can be defined as follows: a set of expert users bookmark scientific papers $\{P_1,\ldots,P_n\} \in \mathcal{P}$. A ranked list of tags $(t_1^j,\ldots,t_{m_j}^j)$ is initially built for each paper P_j by selecting tags from the ontology concepts ($t_i^j \in \mathcal{T}\ \forall\ i,j$). This list is curated *a posteriori* by the expert users. We write T_{rel}^j to denote the set of relevant tags chosen by the experts for paper P_j. The other tags are defined as irrelevant: $T_{\overline{rel}}^j \equiv \mathcal{T} \setminus T_{rel}^j$.

3.2 Tag Disambiguation

The second problem we tackle is tag disambiguation. Since the same literal can appear in the label of several concepts, it is often difficult to disambiguate isolated terms appearing in a paper. For instance, if *anomaly* appears in the text of a scientific paper, should it be related to the *quantum anomaly* concept, to *experimental anomaly* or to *reactor neutrino anomaly*? All are valid scientific concepts but are however very different semantically. Similarly, depending on the context the abbreviation *DM* can mean *Dark matter* (cosmology), *Distance measure* (astronomy), or *Density matrix* (statistical mechanics).

The goal of this second task is to detect such cases and to develop methods to effectively predict which concept an isolated literal should be related to. Obviously, this second task directly relates to our first task, since disambiguating tags produces more

relevant results and hence improves the quality of tag recommendation in the end. Formally, given a term (literal) τ appearing in the text of a paper and a set of automatically selected tags $\{t_1, \ldots, t_m\}$ corresponding to concepts whose label all contain the literal τ, our goal is to automatically select the right tag(s) $t \in T^{\tau}_{rel}$ corresponding to the correct semantics of the literal as chosen by our expert users.

4 Experimental Setting

4.1 Hypotheses

We consider the following hypotheses for the tag recommendation task: i) concepts appearing in the title and the abstract of a paper are highly relevant to that paper, ii) excluding concepts that are too generic yields better recommendations, and iii) using the structure of the ontology can help us recommend better tags. To evaluate those hypotheses, we compare eight different techniques in Section 5.1.

For the tag disambiguation task, we study whether applying clustering techniques on the papers using their concepts as features allows us to disambiguate concepts with a high accuracy. To evaluate this hypothesis, we test two clustering techniques (LDA and K-means) in Section 5.2.

4.2 Metrics

We evaluate the effectiveness of our approach using four standard metrics borrowed from Information Retrieval:

Precision@k defined as the ratio between the number of relevant tags taken from the top-k recommended tags for paper P_j and the number k of tags considered: $P@k = \frac{\sum_{i=1}^{k} \mathbb{1}(t_i^j \in T_{rel}^j)}{k}$ (where $\mathbb{1}(cond)$ is an indicator function equal to 1 when *cond* is true and 0 otherwise).

Recall@k defined as the ratio between the number of relevant tags in the top-k for paper P_j and the total number of relevant tags: $R@k = \frac{\sum_{i=1}^{k} \mathbb{1}(t_i^j \in T_{rel}^j)}{|T_{rel}^j|}$

R-Precision defined as Precision@R, where R is the total number of relevant tags for paper P_j: $RP = P@|T_{rel}^j|$.

Average Precision defined as the average of Precision@k values calculated at each rank where a relevant tag is retrieved over the total number of relevant tags: $AP = \frac{\sum_{i=1}^{|T_{rel}^j|} P@i \; \mathbb{1}(t_i^j \in T_{rel}^j)}{|T_{rel}^j|}$.

Those definitions are valid for one paper only. In the following, we also report values averaged over the entire document collection, e.g., Mean Average Precision (MAP) defined as: $MAP = \frac{1}{n} \sum_{j=1}^{n} AP_j$. The metrics for tag disambiguation are derived similarly (see below Section 5.2).

4.3 Data Sets

We use real data as available on our platform for all our experiments. Our document collection contains all the articles bookmarked by our top-5 most prolific users (user ids 14, 16, 17, 21 and 40). This represents 16'725 scientific papers and 15'083 tags representing 2'157 distinct scientific concepts (out of the 16'725 total number of concepts currently available in our field-specific ontology). If the same paper is bookmarked by more than one user, we take the tags *union* as the relevant set of tags. For the tag disambiguation experiments, we based our experiments on 2'400 articles originating from 6 different top-categories or `ArXiv.org` (400 articles per category).

The experimental data as well as the main scripts we used for our experiments are available on `http://sciencewise.info/media/iswc/`. The data can also be queried using our SPARQL endpoint[8] or browsed online (e.g., `http://data.sciencewise.info/page/bookmarks/2100` gives the bookmark data for paper id 2100).

5 Experimental Results

We report below on our techniques and experimental results for tag recommendation and tag disambiguation.

5.1 Recommending Tags

We compare eight different techniques for tag recommendation below. Most of our approaches are based on term-weighting [15], which is a key technique used in most large-scale information retrieval systems. Basic term-weighting works as follow in our ontology-based context. First, we create an index from the labels of all scientific concepts appearing in the ScienceWISE ontology by considering their stem using Porter's suffix stripping [12]. Then, for each new bookmarked paper, we analyze all the terms appearing in the paper. Given the importance of acronyms in scientific papers, we first determine whether the term is an acronym or not by inspecting its length, capitalization, and by trying to match it to known terms[9]. Two cases can occur at this point: i) if the term is an acronym we consider it *as is* and try to match it to our concept index ii) otherwise, the term is stemmed and then matched using an efficient exact string matching method [9] to the concept index.

We give a brief description of the various methods we experimented with below. We note that each of the following methods was carefully examined and optimized to yield the best possible results we could get after batteries of tests (e.g., we use fined-grained document frequencies and optimal thresholds for all the methods below).

`tf`: Our first approach simply ranks potential tags by counting the number of matches between the terms appearing in the paper and the concept index. While basic, this approach performs relatively well in our context since we consider a restricted

[8] `http://d2r.sciencewise.info/openrdf-sesame/repositories/SW`

[9] We consider that the term is an acronym if it is ≤ 5 letters, all capitalized, and if we cannot find it in the Ubuntu corpus of American words [`http://packages.ubuntu.com/lucid/wamerican`]

number of terms only (our matching process is *mediated* through the ontology). In a standard setting without a field-specific ontology, this approach would perform poorly[10].

tfidf: This second method extends the approach above by applying standard TF*IDF [14]. We use a fine-grained document frequency in this case, based on the top categories of papers in `ArXiv.org` rather than the entire document collection (i.e., IDF is computed based on the paper that share the same `ArXiv.org` topic as the paper being bookmarked), as this performs better in practice.

tf_simpleIDF: In the ScienceWISE ontology, some scientific concepts are marked as "basic". While legitimate, those science concepts are deemed rather general by our users and non-specific to any domain (*mass*, or *velocity* are two examples of such concepts). Under the `simpleIDF` scheme, IDF is not computed; rather, the system simply penalizes basic concepts and systematically puts them at the bottom of the ranked list (i.e., the ranked list of basic tags appears after the ranked list of other tags).

tfidf_title: The scientific terms that appear in titles and abstracts of the scientific papers often carry some special significance. Hence, we modify the TF-IDF ranking to promote the concepts appearing in the title into the top positions of the ranking list. Along similar lines, any concept appearing in the abstract has its TF score doubled (which also promotes it higher up in the list of "suggested tags").

tf_title: The same as above, but discarding IDF and only taking into account TF when ranking.

combined: In this approach we combine `tfidf_title` but use `simpleIDF` to compute the document frequency. As we will see below, only marginally impacts on the effectiveness of the approach while drastically reducing computational complexity for large collections of papers. This is the ranking method that we have decided to deploy on our current production version of ScienceWISE.

ont-depth: Scientific concepts are often organized hierarchically in our ontology, with more specific, sub-concepts deriving from higher-level more general concepts. In this approach, we try to penalize more general concepts (that have a smaller depth in the ontology) and favor more specific concepts. More specifically, we penalize more generic concept by $c\_depth/distance\_from\_root\_concept$ where c_depth is a constant (we use $c\_depth = 1$ below, which yields the best results in our setting).

ont-neighbor: Many scientific concepts are linked to further, related concepts in our ontology. Hence, we take advantage of the semantic graph relating the concepts by improving the scores of those concepts that are direct neighbors of top-k ranked concepts. More specifically, we bump the ranking of direct neighbors of top-ranked concepts by $+c\_neighbor$ (we use $c\_neighbor = 3$ below, which yields the best results in our setting).

Figure 1 compares our different approaches on a Precision VS Recall graph along with the overall results in terms of MAP and R-precision. Results for Precision@k are depicted on Figure 2.

[10] it would lead to a MAP smaller than 1% in our case

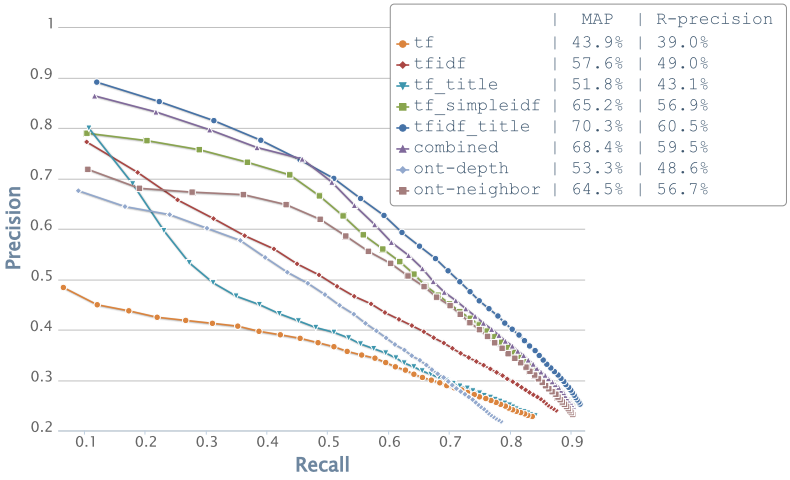

Figure 1: Precision VS Recall for our various tag recommendation approaches

We observe the following:

1. Simple TF ranking yields the worst precision. However, a relatively minor improvement (boosting rank of concepts that occur in the title and abstract, technique called `tf_title` in this paper) greatly improves performance for low k.

2. Performance of the `tfidf_title` is only marginally better than `combined`, with the latter one also being considerably faster (since the global IDF measure does not have to be computed). Both significantly outperform the standard `tfidf` ranking, which demonstrates that one can leverage the structure of scientific texts (where terms in the title and abstract are often very carefully chosen) in order to extract meaningful information.

3. The method leveraging the subsumption relations (*ont-depth*) performs surprisingly poorly. Further variants leveraging the subsumption hierarchies we experimented with behaved even worse. Choosing the right level in the hierarchy seems to be key, and hence favoring too specific (or, conversely, too generic) concepts yields suboptimal results (that are either too specific, and thus unrelated to the paper being analyzed, or too generic and thus are deemed less relevant also).

4. The method based on concept neighborhood (*ont-neighbor*) performs relatively well but is not better than simpler methods. The problem in that case seems to lie in the semantics of the relations between the concepts, which are often arbitrary in our ScienceWISE ontology and hence interconnect semantically heterogeneous concepts. One way of correcting this would be to (automatically or manually) create additional *same-as* or *see-also* relationships in our ontology, and to leverage such relationships to return additional relevant results (we successfully applied such techniques recently on the LOD graph, see [16]).

In summary, the careful use of some specific properties of the ontology (e.g., *basic* concepts) together with information about position of the terms in the document (e.g., in the title or abstract) allow to significantly increase precision in comparison with the baseline methods (increasing MAP up to 70%).

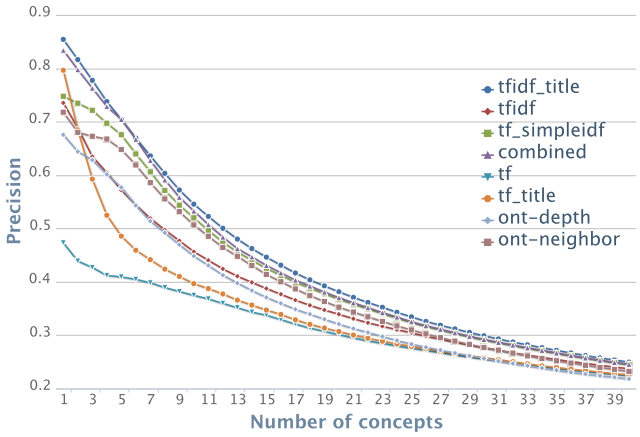

Figure 2: Precision@k of our various ranking techniques for tag recommendation

5.2 Disambiguating Tags

In order to tackle our second problem, we have implemented a special interface, that permits a user to confirm or provide a disambiguation for abbreviations or ambiguous concepts when bookmarking a paper. To help the user in this task, we cluster the collection of bookmarked papers into *topics* in an attempt to guess the correct disambiguation. We start by experimenting with the following techniques:

`lda`: Dirichlet Allocation (LDA) [5] is a standard tool in probabilistic topic modeling. Applied to IR, LDA basically considers that each document is a mixture of a small number of topics and that each word is attributable to one of the topics. It is conceptually similar to probabilistic latent semantic analysis, except that in LDA the topic distributions are assumed to have Dirichlet priors, which often lead to better results in practice. We have use the LDA implementation as available from the Mallet package[11] in our experiments.

`k-means`: works similarly but takes advantage of the well-known k-means clustering technique to cluster the documents.

Since the results produced by both clustering methods only define attribution of each paper to the cluster and does not tell exactly

We consider our data set comprising papers from several disjoint `ArXiv.org` subject classes[12] and split these collections into clusters using LDA and K-Means algorithms. The number of clusters is chosen to be equal to those of primary `ArXiv.org` subject classes.

Next, we use the resulting classification to generate a set of suggestions for the concepts/abbreviation disambiguation. Using our test collection, we determine for each

[11] `http://mallet.cs.umass.edu/`

[12] Each paper on `ArXiv.org` belongs to one or several *Subject classes*, chosen by the authors of the paper

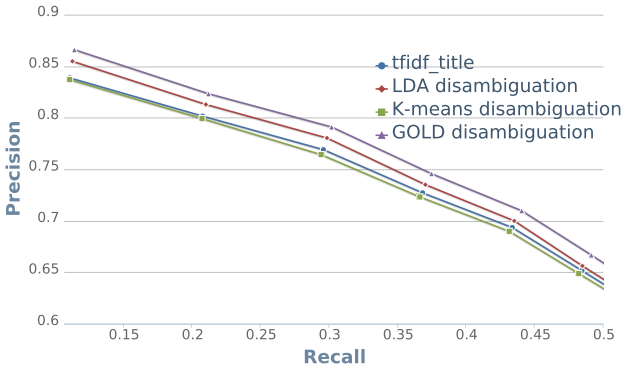

Figure 3: Precision VS Recall using tag disambiguation

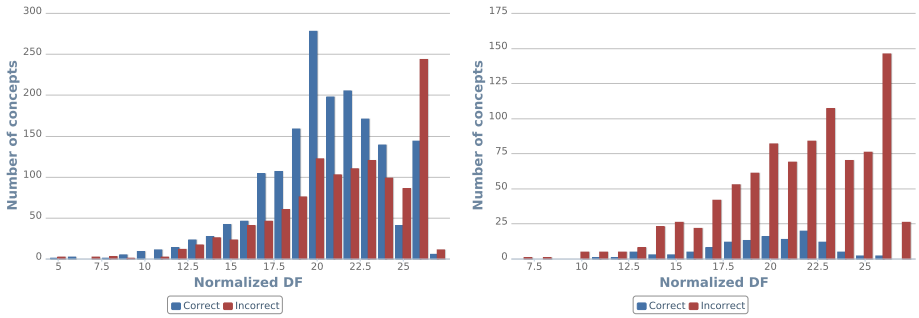

Figure 4: Comparison of document frequency distribution for one-word concepts from the first 5 positions in the ranking (left panel) and from the positions (6–12). NormalizedDF is defined via Eq. (1) in the text.

paper its primary subject class (equivalently, topic) and generated a list of suggestions based on this. The results are shown in Figure 3.

The actual accuracy of LDA-based disambiguation is impressive (**75%**). One can in addition add ontological information to improve the disambiguation process and further boost the accuracy. For example, if among the concepts to disambiguate there is both a concept and subconcept (e.g. *power spectrum* and *matter power spectrum*) and if we provide the most specific concept, the accuracy raises to **88%**. We compare this to the standard k-means clustering algorithm, which only yields an accuracy of **47%**.

Composite Concepts Another approach to the disambiguation problem we experimented with is based on mereology and *composite concepts*. Concepts in a scientific ontology can often be expressed as *composites* of some other ontological concepts. For example, a concept *mass of particle* is a composite of two basic scientific concepts:

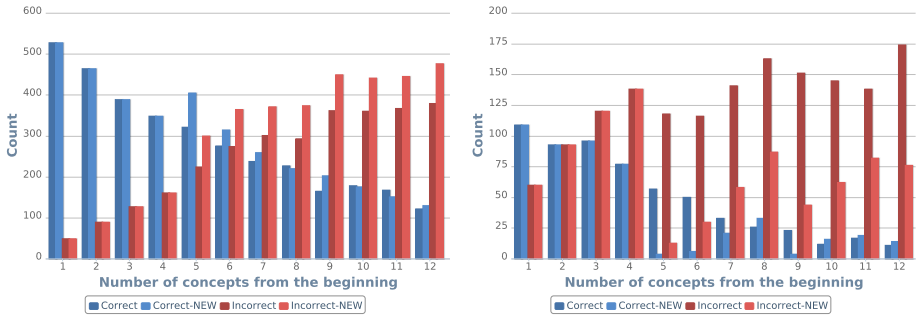

Figure 5: Comparison of acceptance/rejection rate as a function of position in the ranking list before and after penalization of one-word concepts. Left panel shows change of the rejection rate for all concepts, right panel demonstrates rejection rate for one-word concepts.

mass and *particle*. Very often the composite concepts are presented in many different literal forms. Moreover, it is custom to "shorten" the term (e.g. use *mass* instead of *mass of a star*, or simply *cluster* instead of *galaxy cluster*). Although this situation is formally similar to the previous one, it is impossible to guess what concepts should be disambiguated.

We have tested a hypothesis that *one-word concepts more often have a "generic meaning" than their many-words counterparts.*. If this is really the case, a proper tuning of the IDF function would be able to improve the ranking significantly. To determine whether this is indeed the case, we considered the *document frequency* (DF) distribution for the one-word tags. The normalized DF on the x-axis is defined as

$$\text{normalized DF} = \log_{1.5}\left(\frac{\text{number of docs. containing a concept}}{\text{total number of docs. in collection}} \times 10^5\right) \qquad (1)$$

The corresponding histograms are shown in Fig. 4 where one can see (quite surprisingly) that the DF distribution for "correct" and "incorrect" concepts are roughly the same (although the correct ones are shifted somewhat to the lower DF region). Therefore, the one-word concepts bear no clear correlation with the document frequency. Based on these results, we decided to implement a simple strategy for one-word concepts that appear in position 6 and below in our `tf_baseline` ranking list are further penalized. The results of this experiment are shown in Fig. 5. Applied on our tag recommendation strategy, such a disambiguation approach yields and improvement in MAP of about 0.5% on average.

6 Conclusions

In this paper, we addressed the problem of ontology-based tagging of scientific papers. We compared the effectiveness of various methods to recommend and disambiguate tags within a large-scale information system. Compared to classic tag recommendation, the proposed techniques select tags directly from a collaborative, user-driven ontology.

Extensive experiments have shown that the use of a community-authored ontology together with information about the position of the concepts in the documents allows to significantly increase precision over standard methods. Also, several more specific techniques such as ontology-based neighborhood selection, LDA classification and one-word-concept penalization for tag disambiguation yield surprisingly good results and collectively represent a good basis for further experimentation and optimizations.

References

1. Karl Aberer, Alexey Boyarsky, Philippe Cudré-Mauroux, Gianluca Demartini, and Oleg Ruchayskiy. An integrated socio-technical crowdsourcing platform for accelerating returns in escience. In *ISWC (Outrageous Ideas Track)*, 2011.
2. Karl Aberer, Alexey Boyarsky, Philippe Cudré-Mauroux, Gianluca Demartini, and Oleg Ruchayskiy. ScienceWISE : a Web-based Interactive Semantic Platform for scientic collaboration. In *ISWC (Demonstration Track)*, 2011.
3. Benjamin Adrian, Leo Sauermann, and Thomas Roth-berghofer. Contag: A semantic tag recommendation system. In *Proceedings of ISemantics 07*, pages 297–304. JUCS, 2007.
4. Pierpaolo Basile, Marco Degemmis, Anna Lisa Gentile, Pasquale Lops, and Giovanni Semeraro. The jigsaw algorithm for word sense disambiguation and semantic indexing of documents. In Roberto Basili and Maria Teresa Pazienza, editors, *AI*IA*, volume 4733 of *Lecture Notes in Computer Science*, pages 314–325. Springer, 2007.
5. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
6. Adriana Budura, Sebastian Michel, Philippe Cudré-Mauroux, and Karl Aberer. Neighborhood-based tag prediction. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvnen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 608–622. 2009.
7. Christiane Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.
8. Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in social bookmarking systems. *AI Commun.*, 21(4):231–247, 2008.
9. Donald E Knuth, Jr James H Morris, and Vaughan R Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977.
10. Marek Lipczak. Tag recommendation for folksonomies oriented towards individual users. *ECML PKDD Discovery Challenge*, 2008.
11. Gilad Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin, editors, *WWW*, pages 953–954. ACM, 2006.
12. M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
13. S.S. Sahoo, A. Sheth, and C. Henson. Semantic provenance for escience: Managing the deluge of scientific data. *Internet Computing, IEEE*, 12(4):46–54, 2008.
14. G. Salton and M.J. McGill. Introduction to modern information retrieval. 1986.
15. Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, August 1988.
16. Alberto Tonon, Gianluca Demartini, and Philippe Cudre-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *SIGIR*, 2012.
17. Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*, 2006.