On Tag Recommendation for Expertise Profiling: A Case Study in the Scientific Domain

Isac S. Ribeiro, Rodrygo L. T. Santos,
Marcos A. Gonçalves, Alberto H. F. Laender
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
{isacsandin,rodrygo,mgoncalv,laender}@dcc.ufmg.br

ABSTRACT

Building expertise profiles is a crucial step towards identifying experts in different knowledge areas. However, summarizing the topics of expertise of a given individual is a challenging task, primarily due to the semi-structured and heterogeneous nature of the documentary evidence available for this task. In this paper, we investigate the suitability of tag recommendation as a mechanism to produce effective expertise profiles. In particular, we perform a large-scale user study with academic experts from different knowledge areas to assess the effectiveness of multiple supervised and unsupervised tag recommendation approaches as well as multiple sources of textual evidence. Our analysis reveals that traditional content-based tag recommenders perform well at identifying expertise-oriented tags, with article keywords being a particularly effective source of evidence across profiles in different knowledge areas and with various levels of sparsity. Moreover, by combining multiple recommenders and sources of evidence as learning signals, we further demonstrate the effectiveness of tag recommendation for expertise profiling.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*User profiles and alert services*

General Terms

Experimentation, Measurement

Keywords

Expertise profiling; tag recommendation; learning to rank

1. INTRODUCTION

The online possibilities for people to interconnect and share experiences and knowledge is unprecedented. One domain that can benefit tremendously from such interconnection possibilities is science. In particular, researchers from multidisciplinary scientific areas can establish new contacts, form networks, and collaborate on challenging research problems. However, the large number of researchers acting online and searching for relevant contacts and partnerships requires new tools and services for automatically finding experts in a given knowledge area. Two challenges need to be addressed for these services to work properly: (1) the construction of *profiles* that can meaningfully describe a researcher's expertise and (2) the design of algorithms capable of ranking researchers according to such expertise profiles. The latter task, denoted *expert finding*, has recently received a lot of attention from the academic community [3]. The task of *expertise profiling*, on the other hand, has received considerably less attention [27, 28] and is hence our focus.

There are several challenges related to the construction of meaningful expertise profiles, such as: (i) balancing between conciseness versus representativeness, (ii) identifying the best topics to summarize sometimes very long careers, (iii) dealing with the evolution of topics of interest over time, and (iv) extracting and aggregating several sources of expertise evidence including online curricula vitae, pages in different digital libraries, and social networks. To address some of these challenges, we approach the expertise profiling problem as a problem of *tag recommendation for people* (or *people tagging*). Although tag recommendation has been studied before for a range of different types of online media [10, 15, 18, 19, 30], people tagging is an area which has been less investigated and with opportunities for improvements.

In this paper, we investigate the suitability of tag recommendation for producing effective expertise profiles. To this end, we perform a large-scale user study involving 1,288 respondents (of 5,355 contacted) among the most prominent researchers from different areas of knowledge in Brazil to assess the effectiveness of people tagging for expertise profiling. As a source of expertise evidence, we leverage each researcher's curriculum vitae (CV) as archived in the Lattes Platform,[1] an internationally renowned initiative that manages information about science, technology, and innovation related to individual researchers and research institutions in Brazil [20]. Our investigation contrasts three representative content-based tag recommendation algorithms from the literature, exploiting three textual sources of expertise evidence from the publications listed in each researcher's Lattes CV, namely, their title, abstract, and keywords. Our experimental results demonstrate the effectiveness, completeness, and robustness of the expertise profiles built through tag recom-

[1] `http://lattes.cnpq.br`

WSDM'15, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2685320.

mendation for researchers in different areas of knowledge and for various levels of sparsity of the available textual evidence. While keyword-based expertise profiles are shown to be particularly effective, we further investigate the suitability of multiple learning to rank (L2R) strategies for combining the considered recommendation algorithms and sources of evidence as learning features. Our results show that the best learning strategies, which are based on an ensemble of rankers, can produce results about 21.5% better than the best tag recommender baseline used in isolation. Moreover, we show that a profiling effectiveness of 80% of that attained by the most effective learned model can be achieved using only two features, which further hints on the potential applicability of tag recommendation in real deployments. To the best of our knowledge, the size of this investigation in terms of coverage (a whole country), disparate areas of knowledge, algorithms, and data source alternatives is unique.

In summary, the major contributions of this paper are:

1. A large-scale user study for expertise profiling with the most prominent Brazilian researchers in different knowledge areas and with various career lengths.

2. A thorough empirical assessment of the effectiveness of multiple unsupervised and supervised tag recommenders for automatic expertise profiling.

In the remainder of this paper, Section 2 covers related work in several areas such as expertise profiling, tag recommendation, and evaluation of tag recommenders. Section 3 details the methodology underlying our experimental evaluation, including the procedures for mining and ranking candidate tags, as well as for collecting relevance judgements in our large-scale user study. Section 4 discusses the results of our experimental evaluation. Finally, Section 5 provides our concluding remarks and directions for future work.

2. RELATED WORK

Expertise profiling is an essential component of any expert finding or expert retrieval system [3]. There has been some investigation on the theme for a while (see, for instance, [2] and [12]), but the interest remains, with very recent work dedicated to the temporal aspects of expertise [14, 27]. In this work, we approach the expertise profiling problem as one of *people tagging* [28]. This solution has several advantages, such as: (i) uniform and concise representation of the profile, (ii) existence of previous methods in other domains to suggest or rank a set of candidate tags, (iii) capability of generating different "tag clouds" for different periods in a person's career, (iv) uniform treatment of the sources of information as "bags of candidate tags", etc. These advantages can help a people-tagging approach overcome some of the challenges in expertise profiling mentioned in Section 1. However, people tagging is a problem much less investigated than tag recommendation for other types of "objects".

Tag recommendation approaches have been proposed for a wide range of media types, mainly in the so-called Web 2.0 [4, 5, 6, 10, 15, 18, 19, 30]. The state-of-the-art techniques for this task exploit co-occurrence patterns with previously assigned tags, expanding an initial set of tags I_0 for an object o with other tags that co-occur with tags in I_0, although in different objects of the collection. The most relevant tags can then be used to induce a tag cloud to represent the object o. For instance, Canuto et al. [10] compared several L2R algorithms for the task of tag recommendation, but they do assume that I_0 is not empty. When there are no tags initially available, as is the case in many scenarios for people tagging, such methods do not work properly [23]. Other works exploit connections among objects that enable the propagation of tags from a tagged object to an untagged one [21, 29, 32, 38]. This can be seen as a type of collaborative filtering approach [13]. In contrast, in this paper, we adopt a pure content-based approach, by exploiting only generally available evidence from the contents of the publications written by the researcher to be tagged.

A relevant approach to our work, yet in the context of object tagging, was proposed by Venetis et al. [33]. In particular, they defined a series of metrics that capture structural properties of a tag cloud. For instance, the coverage of a tag cloud represents the fraction of all objects that can be retrieved by the tags in the cloud. Based on the proposed metrics, they developed a satisfaction model to evaluate the quality of a tag cloud for a given search task. This model considers the probability that the tags in the cloud will fail to satisfy the information need of a user who employs them to browse or search a document collection. Using this model, they performed a quantitative analysis of several algorithms for tag selection. The best performing among these algorithms are used as alternative tag recommenders in our analysis.

The closest approach to ours is the work of Serdyukov et al. [28], who proposed a method for people tagging for expertise profiling in an enterprise domain. While our primary goal is to assess the suitability of tag recommendation for expertise profiling, there are important differences between their work and ours. First, our domain is a scientific one, meaning that the evidence we exploit is completely different. Indeed, while they leveraged features such as web documents, discussion lists and enterprise search clickthrough, we exploit features that are specific to scientific documents, such as title, abstract, and keywords. Second, in their evaluation, they contrasted the recommended tags with self-created profiles. If a recommended tag was not in the profile created by the own employees, the tag would be automatically considered as irrelevant. Such an arguably strict form of evaluation may explain the low performance figures reported in their work. In contrast, we employ a TREC-like pooling approach to gather relevance judgements for tags suggested by nine different recommenders (i.e., three tag recommendation algorithms deployed with three different sources of evidence). Finally, they approached tag recommendation as a classification task, using a logistic regression classifier to determine whether a tag was relevant or not, with the confidence of this classification used to induce an overall ranking. On the other hand, we approach this task as an explicit L2R task, and contrast nine state-of-the-art L2R algorithms to leverage the scores produced by our nine considered tag recommenders as learning features.

Finally, the evaluation of tag recommenders is a research issue by itself. Most previous works rely on an automatic evaluation procedure in which part (usually 50%) of the tags assigned to an object already in the system are used for training purposes, while the remaining ones are used as the gold standard that should be predicted by the recommender. This is mostly due to the inherent difficulties and cost associated with manual user evaluations. Moreover, it is arguably hard for an assessor to judge if a tag assigned by a different user (i.e., a user who has not uploaded the object

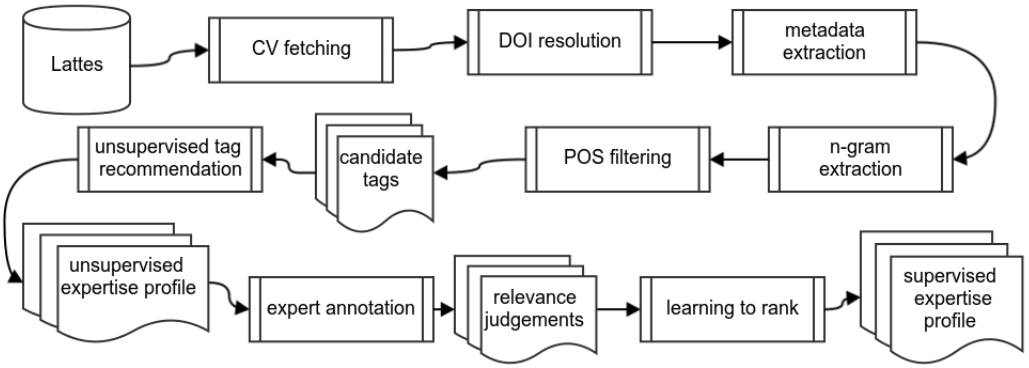


Figure 1: Overview of our experimental methodology.

or is not necessarily familiar with it) is relevant to an object whose interpretation may be subjective (e.g., an image or a video). The problem with this type of evaluation is that a possibly relevant tag for an object may be recommended but it will not be considered as relevant if it is not in the current gold standard. Due to the aforementioned difficulties, only a handful of works (e.g., [7, 25, 29, 31, 34]) assessed tag recommendations with manual user evaluations, usually performed in a very small scale. In contrast, we present a large-scale study involving 1,288 respondents (of 5,355 contacted) among the most prominent researchers in different areas of knowledge currently working in Brazil. Moreover, these assessors can be considered as ideal ones, since they are assessing representations of their own scientific production over the years. In this sense, the dataset we build is a valuable asset for other researchers working on related areas. We intend to make this dataset available soon.

3. EXPERIMENTAL METHODOLOGY

Tag recommendation approaches have been traditionally evaluated by partitioning existing sets of tags, contributed by a multitude of different users, into training and test. The evaluation of people tagging poses additional challenges, primarily because the people being tagged must agree on the relevance of the tags assigned to them. The importance of such an approval is further exacerbated when tags have an expertise orientation, which is our focus here. Despite this, carrying out large-scale user evaluations is often costly and time consuming, which may explain the scarcity of such evaluations in the literature. In this section, we describe the experimental methodology adopted in our work, which supports one such evaluation. In particular, we discuss the selection of a representative target set of experts, the acquisition of expertise evidence about them, the generation of candidate expertise profiles through tag recommendation, and the assessment of the generated profiles. Figure 1 provides an overview of the entire methodology. Statistics of the produced test collection are presented afterwards.

3.1 Data Acquisition

The first step towards producing a test collection for expertise profiling is to choose a representative set of experts. In 2008, the Brazilian National Council for Scientific and Technological Development (CNPq) introduced a funding program to foster networks of collaborative research in several areas considered strategic for the country. Together, the 123 awarded research groups, called National Institutes of Science and Technology (INCTs),[2] comprise over 6,000 of the most prominent researchers from all areas of knowledge working in Brazil. Our test collection is built around these researchers as a representative set of experts to be profiled. In particular, we believe that the breadth of their expertise and their heterogeneous career paths poses realistic challenges to evaluating expertise profiling approaches.

Having selected a set of experts for evaluation, we must collect evidence of their expertise, to be mined in order to produce effective expertise profiles. To this end, as discussed in Section 1, we resorted to the Lattes Platform, a publicly accessible repository of academic information maintained by CNPq, which archives up-to-date CVs for researchers working in both public as well private research institutions in Brazil. In particular, of over 6,000 researchers in our target group, we managed to collect the Lattes CV of 5,355 of them.[3] From each collected CV, we extracted information about all publications up to April 2014. To establish a common ground for researchers in different areas, we focused on journal publications, which are generally seen as the main venue for research dissemination in most areas, and leave the exploration of other sources of evidence for future work.

Having collected the CVs and extracted the titles of all journal publications contained in them, we started to crawl additional metadata corresponding to each publication. For the publications with no Digital Object Identifier (DOI), we performed a query using their citation data to the API of the CrossRef service.[4] With the provided and discovered DOIs, the next step was to collect the metadata using the services of the respective publishers. In particular, we restricted ourselves to the twenty most prolific publishers, which respond for more than 80% of the total number of publications to be collected, as shown in Figure 2. For each publication from these publishers, we extracted its abstract and list of keywords. In addition to the publication title already extracted from the Lattes CV, these metadata form three sources of textual evidence for expertise mining.

[2] `http://goo.gl/FdjqRo`
[3] The Lattes CV for the remaining researchers could not be collected due to persistent download failures.
[4] `http://www.crossref.org`

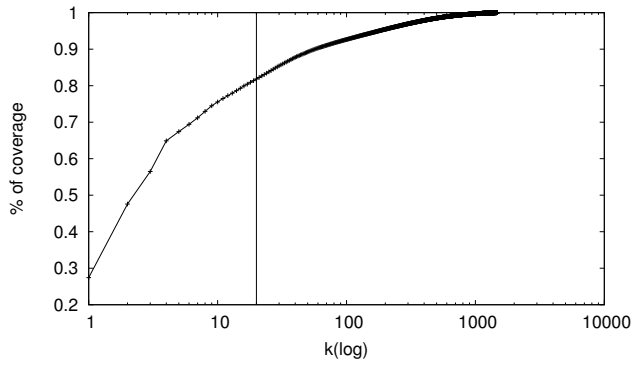


Figure 2: Publications covered by the top k publishers.

3.2 Profile Generation

After acquiring publication metadata to construct different sources of expertise evidence, namely, title, abstract, and keywords, the next step was to generate a candidate expertise profile for each researcher. To this end, we mined candidate tags from each source of evidence by extracting n-grams with varying sizes (from $n = 1$ to $n = 3$). In order to further discard tags unlikely to represent any topic of expertise, we performed a small pilot study with volunteers, including graduate students and staff from our research group. Based upon a random sample with 7,123 tags, this pilot study identified 57 part-of-speech (POS) patterns [22] commonly associated with syntactically malformed tags. In particular, we selected POS patterns with minimum confidence of 90% and minimum support of 15 tags to maximize F1 in our pilot study. Accordingly, the selected POS patterns were used to filter out potentially malformed tags from our candidate set. An example POS filter is NN-IN-VBG (i.e., noun, preposition, gerund), which removed tags such as "proposal after executing" and "testing for validating".

Finally, in order to avoid the limitations of self-created ground-truth profiles, as discussed in Section 2, we generated a pool of diverse tags to be assessed by each researcher. To this end, we chose three representative content-based tag recommendation algorithms, which are commonly used as ranking components within state-of-the-art machine-learned tag recommenders from the literature [33]:

POP: A popularity-based tag recommendation algorithm, which returns the top k tags according to the term frequency (TF) of each tag.

TFIDF: A traditional content-based tag recommendation algorithm, which chooses the top k tags according to their TF×IDF score, where IDF stands for the inverse document frequency of a tag, which is based on the number of publications where it occurs. This algorithm is similar to POP, with the exception that the IDF component further helps demote too popular tags in favor of more discriminative ones.

COV: A recommendation algorithm that seeks to maximize the coverage of a researcher's publications that can be retrieved by the top k selected tags. This algorithm uses an iterative, greedy strategy for choosing the top k tags by selecting, at each iteration, the tag that covers the most yet uncovered publications.

Each of these three algorithms was applied on top of each of the three considered sources of evidence, namely, title, abstract, and keywords, hence producing nine different recommenders. Lastly, we pooled the ranking produced by each of these recommenders to depth 50 in order to produce a final pool of tags to be assessed by each researcher.

3.3 Profile Assessment

The next step in our evaluation methodology was to gather relevance assessments for each of the tags pooled for each researcher using the nine devised recommenders. To this end, we invited the 5,355 researchers for which we could produce a candidate expertise profile to assess the relevance of the tags in this profile. Of the 5,355 contacted researchers, 1,288 responded to our invitation and participated in the assessment in a period of two weeks in July 2014. Such a high response rate of nearly 25% testifies to the relevance of the study for the community itself. Figure 3 provides a breakdown of the number of respondents per area of knowledge. From the figure, we note that the response rate was roughly consistent across areas, with Health Science—the largest community in our study—contributing most of the respondents.

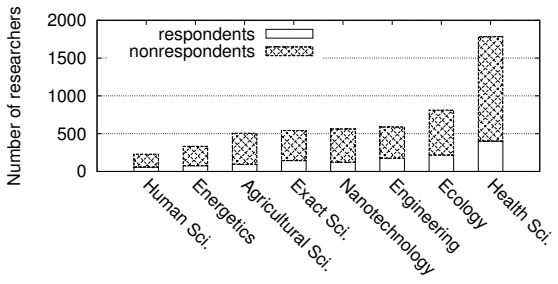


Figure 3: Number of respondents per area of knowledge.

A screenshot of the relevance assessment interface (translated from Portuguese) is presented in Figure 4. During the assessment, each researcher was presented with 60 tags in no particular order, selected in a round-robin fashion from the top 50 tags returned by each of the nine pooled recommenders. The researcher was then asked to assess the relevance of each tag according to the following four-point scale:

1. *"The tag is malformed (a spurious tag)"*
2. *"The tag is well-formed, but is not relevant to describe my work"*
3. *"The tag is well-formed, but is only partially relevant to describe my work"*
4. *"The tag is well-formed and is highly relevant to describe my work"*

Figure 5 shows a distribution of assessed tags for each of the eight areas of knowledge considered in our study. From the figure, we first note that the distribution of tags assessed according to the aforementioned four-point scale is consistent across the different areas. More importantly, we note that, for all knowledge areas, the proportion of malformed tags is lower than that of the other classes taken together. This further demonstrates the appropriateness of the chosen recommenders for identifying potentially relevant tags.

Dear Rodrygo Luis Teodoro Santos

As part of a research project of the National Institute of Science and Technology for the Web, we have developed new methods to generate a representative list of tags to describe the topics of expertise of researchers based on their scientific production. We would like to invite you to validate this list of topics for your particular case.

The tags listed below were automatically generated based on the publications available in our Lattes curriculum with the goal of describing the most representative topics of your research.

For each tag, please indicate one of the following options:

1. The tag **is malformed** (a spurious tag)
2. The tag is well-formed, but **is not relevant** to describe my work
3. The tag is well-formed, but **is only partially relevant** to describe my work
4. The tag is well-formed and **is highly relevant** to describe my work

Tag	Classification	Tag	Classification	Tag	Classification
learning-to-rank	○ 1 ○ 2 ○ 3 ○ 4	missing full text	○ 1 ○ 2 ○ 3 ○ 4	whens and hows	○ 1 ○ 2 ○ 3 ○ 4
web search engines	○ 1 ○ 2 ○ 3 ○ 4	query suggestions	○ 1 ○ 2 ○ 3 ○ 4	component-based software development	○ 1 ○ 2 ○ 3 ○ 4
expert search	○ 1 ○ 2 ○ 3 ○ 4	note that existing	○ 1 ○ 2 ○ 3 ○ 4	search	○ 1 ○ 2 ○ 3 ○ 4
search result diversification	○ 1 ○ 2 ○ 3 ○ 4	wizard tool	○ 1 ○ 2 ○ 3 ○ 4	query-dependent	○ 1 ○ 2 ○ 3 ○ 4
mimicking web search	○ 1 ○ 2 ○ 3 ○ 4	documents with respect	○ 1 ○ 2 ○ 3 ○ 4	web services	○ 1 ○ 2 ○ 3 ○ 4

Figure 4: Screenshot of the tag assessment system (translated from Portuguese).

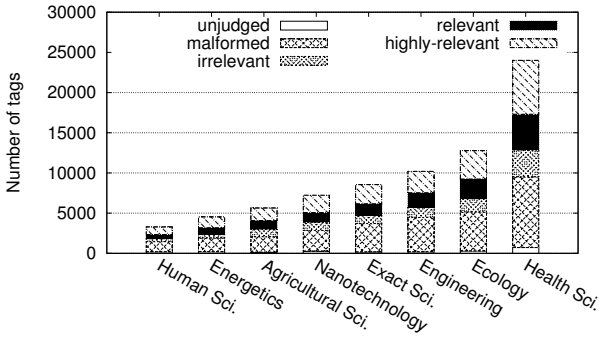


Figure 5: Distribution of assessed tags per area.

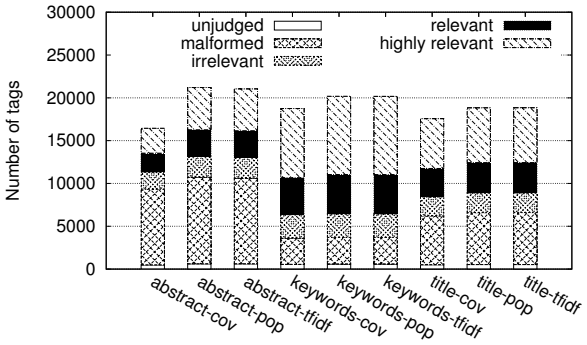


Figure 6: Distribution of assessed tags per recommender.

A similar distribution to the one shown in Figure 5 is shown in Figure 6, however across the nine tag recommenders used for pooling. From the figure, we note the higher incidence of malformed tags among abstract-based tag recommenders, despite the POS filtering step applied to remove such tags. On the other hand, titles have relatively fewer malformed tags, with keywords having the best performance in this aspect. As we will further explore in Section 4, this can be partially explained by the space constraints inherent to each of these sources of evidence and the corresponding care taken by a researcher while filling each of them.

Salient statistics of the produced test collection are summarized in Table 1. From the table, it is worth noting the representativeness of the sample data obtained with respect to the universe of INCT researchers contacted. Equally interesting are the idiosyncrasies of individual areas, particularly in terms of the average number of publications per researcher (#ppr) and the average number of tags the respondents deemed partially (#rtpr) and highly relevant (#htpr) to represent their expertise. As we will discuss in Section 4, this test collection is used to assess the suitability of tag recommendation as a mechanism to automatically generate expertise profiles. In addition, we further leverage the produced relevance assessments as labeled data for learning to rank for expertise profiling. We intend to make the produced test collection available for the benefit of the community.

	contacted		respondents			
area	#res	#ppr	#res	#ppr	#rtpr	#htpr
Agricultural Sci.	507	43.06	94	50.06	11.57	16.71
Ecology	812	25.35	216	27.00	11.20	16.45
Energetics	333	29.26	76	32.22	10.53	17.92
Engineering	588	24.26	175	21.74	10.18	15.32
Exact Sci.	542	49.95	144	56.13	9.85	16.65
Health Sci.	1,783	48.13	403	52.03	11.00	16.68
Human Sci.	229	15.31	58	15.33	8.69	16.36
Nanotechnology	561	50.32	122	58.75	9.93	17.77
Total	5,355	39.41	1,288	41.85	10.60	16.62

Table 1: Statistics of the generated test collection, including the number of researchers (#res) and of publications per researcher (#ppr) for contacted as well as respondent researchers. For the latter, we also show the number of relevant (#rtpr) and highly relevant (#htpr) tags per researcher.

4. EXPERIMENTAL EVALUATION

In this section, we thoroughly analyze the suitability of tag recommendation for expertise profiling. In particular, we aim to answer the following research questions:

Q1. How effective is tag recommendation as a mechanism for automatic expertise profiling?

Q2. How complete is tag recommendation as a mechanism for automatic expertise profiling?

Q3. How robust is tag recommendation as a mechanism for automatic expertise profiling?

Q4. Can we effectively learn to recommend expertise tags?

The results of our analysis are discussed in the subsequent sections, which address each of these questions in turn.

4.1 Profiling Effectiveness

To address question Q1, we assess three representative content-based tag recommendation algorithms from the literature, namely TFIDF, POP, and COV [33]. As discussed in Section 3, each algorithm is provided with textual evidence of a researcher's expertise, derived from the researcher's publication titles, abstracts, or keywords, and produces an expertise profile for the researcher. Figure 7 summarizes the retrieval effectiveness of the expertise profiles generated by different combinations of recommendation algorithm and textual evidence. For each combination, we report normalized discounted cumulative gain (nDCG) figures for multiple ranking cutoffs [1]. In particular, to compute nDCG, we assign partially and highly relevant tags the relevance labels 1 and 2, respectively. Malformed and irrelevant tags are both assigned a relevance label 0. For the sake of readability, error bars denoting 95% confidence intervals are omitted as they are smaller than the plotted symbols.

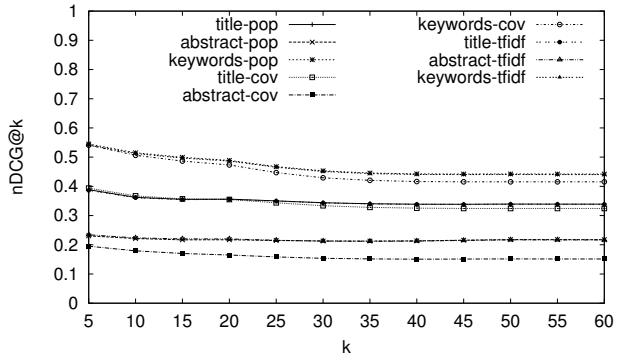


Figure 7: Profiling effectiveness in terms of nDCG at various ranking cutoffs. Error bars are ommited for readability.

Regarding the different tag recommendation algorithms considered in our investigations, from Figure 7, we first note that TFIDF performs similarly to POP, which is somewhat expected. In particular, as shown in Figure 8, nearly 80% of all tags occur only once, and over 90% occur at most twice in the entire corpus. In this scenario, with the majority of tags showing similar scarcity, the IDF component has little impact on the final ranking produced by the TFIDF algorithm. As a result, the ranking produced by TFIDF resembles a pure TF-based ranking, which is essentially equivalent to the ranking produced by the content-based POP algorithm.

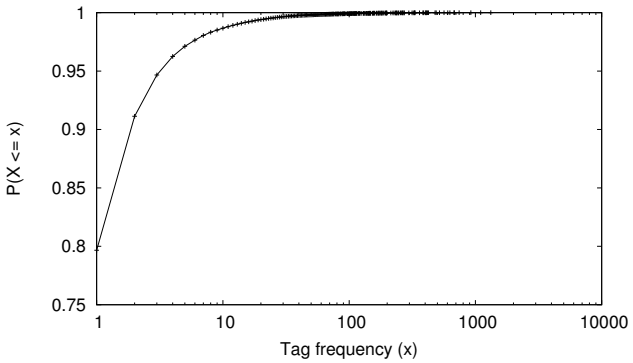


Figure 8: Cumulative distribution of tag frequencies.

From Figure 7, we also note that both TFIDF and POP outperform the COV algorithm. In order to explain this behavior, Table 2 shows the average pairwise overlap among the top 10 tags retrieved by the POP algorithm, in terms of the Jaccard coefficient obtained by comparing the publications where these tags are mentioned. From the table, we observe that the top tags returned by POP have little overlap with respect to one another in terms of the publications that they cover. Indeed, the average pairwise overlap of the tags ranked by POP is as low as 0.102805 in the best case, when considering publication abstracts as a source of expertise evidence. As a result, explicitly seeking to promote a high coverage of these publications, as proposed by the COV algorithm, has little extra benefit, and can even harm the effectiveness of the recommended tags, as is the case when using abstracts. Moreover, COV is computationally more expensive than TFIDF and POP, and hence should be eschewed in favor of these faster approaches, whenever effectiveness is the primary expertise profiling concern.

evidence	Jaccard
abstract	0.102805 ± 0.009586
keywords	0.097010 ± 0.007868
title	0.044094 ± 0.003939

Table 2: Average pairwise Jaccard coefficient for the top 10 tags generated by POP using different textual evidence.

Regarding which textual features are useful expertise indicators, keywords provide a more effective expertise evidence compared to titles, which in turn perform more effectively than abstracts. These observations are consistent across the entire range of nDCG in Figure 7 and suggest that the more constrained the source of evidence, the more useful an expertise indicator it is. Indeed, with limited room for conveying an idea, researchers must strive to ensure that this idea is carefully described. Conversely, less constrained sources of evidence, such as abstracts, provide more freedom for content generation, yet they are more prone to noise. As an illustrative example, Table 3 shows the top 5 tags generated by POP using each of the considered sources of evidence for building an expertise profile for one of the authors of this paper. Looking at the generated tags, we observe that while keyword-based tags are more likely to represent a valid topic

of expertise, such as "web search" and "learning to rank", the other sources of evidence may contribute with further plausible tags, such as "search engines" and "weighting models".

keywords	title	abstract
web search	digital	weighting models
learning to rank	search result	learning to rank
relevance	learning to rank	combination
diversity	component-based	search result
digital libraries	search engines	ambiguous query

Table 3: Top 5 tags generated by POP using keywords, title, and abstract for one of the authors of this paper.

Recalling question Q1, the results in this section attest the suitability of traditional content-based tag recommendation algorithms for identifying relevant topics of expertise. In particular, a recommender based on the POP algorithm with tags extracted from the researchers' publication keywords generates the most effective expertise profiles.

4.2 Profiling Completeness

The results in Section 4.1 show that traditional content-based tag recommendation algorithms can produce effective expertise profiles. A natural question that arises in this scenario is how complete these profiles are. In particular, to address Q2, we analyze the completeness of the expertise profiles produced through tag recommendation. To assess the completeness of an expertise profile, we borrow the tag coverage metric proposed by Venetis et al. [33]. In our case, coverage@k measures the fraction of all publications from a researcher that can be retrieved by the top k tags in this researcher's expertise profile. Figures 9(a)-(c) show such coverage figures at different ranking cutoffs k for all considered combinations of recommendation algorithm and textual evidence. Once again, error bars are omitted for readability.

From Figures 9(a)-(c), we first observe that, with only a few top retrieved tags, most combinations of recommendation algorithm and textual evidence are able to cover the majority of the researchers' scientific production. In particular, abstract and keywords-based profiles have coverage@10 figures of nearly 90%. Title-based expertise profiles, on the other hand, require around 50 tags to attain the same level of coverage. Regarding the three considered recommendation algorithms, COV, which directly aims at improving coverage, is particularly effective at earlier ranking cutoffs when applied on abstracts. On titles and keywords, COV is more effective at deeper cutoffs, with statistically significant improvements compared to both TFIDF and POP.

Recalling question Q2, the results in this section attest the completeness of the expertise profiles built through tag recommendation. Indeed, all considered recommenders are able to comprehensively convey a researcher's expertise with only a few top ranked tags. The COV algorithm, which seeks to promote tags with a high coverage of a researcher's publications, can be particularly advantageous to this end.

4.3 Profiling Robustness

The results presented thus far attest the suitability of tag recommendation for expertise profiling in terms of the effectiveness and completeness of the recommended tags. In this section, we investigate two factors that may affect such an effective performance. In particular, in order to address Q3, we assess the effectiveness of tag recommendation for building expertise profiles in different areas of knowledge, as well as for expertise evidence of different levels of sparsity. Regarding the latter, Figure 10 shows the profiling effectiveness of all considered combinations of tag recommendation algorithm and textual evidence in terms of nDCG@10 for researchers with different numbers of journal publications.

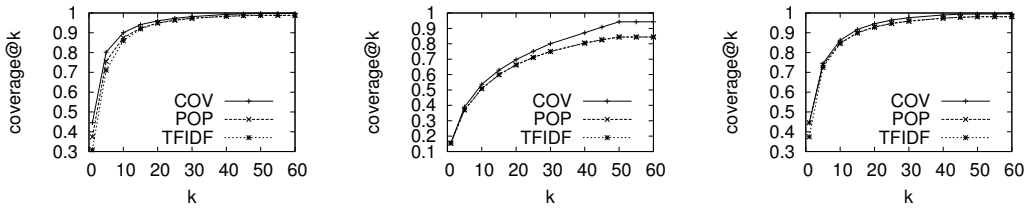
From Figure 10, we first observe that all approaches perform reasonably well even for researchers with only a few publications in their curriculum. This highlights the robustness of these approaches against sparse expertise evidence. Nonetheless, we note that the larger the researcher's curriculum, the more effective the produced expertise profile. Indeed, larger curricula tend to offer both breadth for covering distinct expertise topics, as well as depth for reassuring the importance of core topics. However, larger curricula may also introduce noise in the tag recommendation process. This is particularly the case for more verbose sources of evidence, such as abstracts. In particular, besides achieving an inferior performance compared to recommenders based on titles and keywords, recommenders based on abstracts are also less influenced by the amount of available evidence.

In addition to assessing the impact of the amount of available expertise evidence, Figure 11 shows the profiling effectiveness of the aforementioned combinations of recommendation algorithm and textual evidence for the eight areas of knowledge represented in our corpus, as described in Section 3. From Figure 11, we note three clear patterns of effectiveness across the different areas of knowledge, corresponding to the three different textual evidence employed. In particular, abstracts lead to a more stable effectiveness across areas, whereas keywords and titles are more unstable. While all considered areas have roughly the same number of relevant tags per researcher, as also discussed in Section 3, some interesting results can be observed. For instance, Agricultural Science generally shows the highest performances among all areas. Health Science, which by far has the largest number of relevant tags as well as of researchers in our corpus, has an intermediate performance for all considered recommendation algorithms and textual evidence. Lastly, Human Science has one of the lowest performances with abstracts and titles, but one of the best with keywords. This suggests that different sources of evidence may provide complementary benefits, as we will discuss in the next section.

Recalling question Q3, the results in this section demonstrate the robustness of tag recommendation for expertise profiling. While the availability of larger expertise evidence improves the effectiveness of the produced profiles, the considered combinations of recommendation algorithm and textual evidence perform robustly even for sparse data. Different areas of knowledge, on the other hand, have different impact on tag recommendation, yet without noticeably compromising the attained effectiveness for any particular area.

4.4 Learning Expertise Profiles

The results in Sections 4.1 through 4.3 demonstrate the effectiveness, completeness, and robustness of tag recommendation for expertise profiling, with keyword-based tags being particularly effective. From these results, one may argue that relying on keywords alone could be sufficient for generating an effective expertise profile, given that keywords are generally carefully informed by the researchers themselves. Nevertheless, other sources of evidence could provide further relevant tags that are not mentioned among the re-



(a) Abstract (b) Title (c) Keywords

Figure 9: Coverage of publications for multiple tag recommendation algorithms and textual evidence.

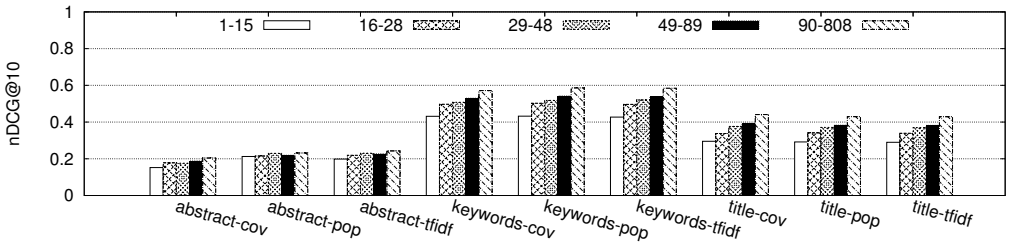


Figure 10: Profiling effectiveness for researchers with different numbers of journal publications.

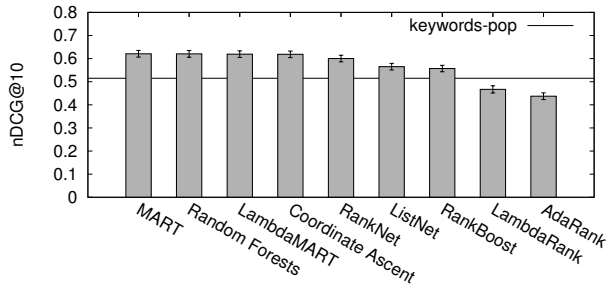


Figure 12: Profiling effectiveness of L2R algorithms.

searchers' publication keywords. In this section, we assess the complementarity of these sources of evidence combined to the TFIDF, POP, and COV recommendation algorithms. To this end, we address question Q4 by seeking to learn an effective ranking model for expertise profiling.

To address question Q4, we assess the suitability of the aforementioned tag recommenders as features for learning to rank for expertise profiling. In particular, we test nine state-of-the-art L2R algorithms from the literature, as implemented in the RankLib library:[5] MART [17], RankNet [9], RankBoost [16], AdaRank [36], Coordinate Ascent [24], LambdaMART [35], LambdaRank [26], ListNet [11], and Random Forests [8]. For each of these L2R algorithms, we perform a 5-fold cross validation in order to optimize nDCG@10, with three folds used for training, one for validation, and one for testing. Accordingly, in Figure 12, we report nDCG@10 figures averaged across the test folds for each individual L2R algorithm. Error bars indicate 95% confidence intervals for the reported means. Finally, a horizontal line indicates the performance of the single most effective feature identified in Figure 7, namely, keywords-pop.

[5] `http://sourceforge.net/p/lemur/wiki/RankLib/`

From Figure 12, we observe a statistical tie among MART, Random Forests, LambdaMART, and Cordinate Ascent. In addition, these four L2R algorithms are statistically superior to the remaining five (RankNet, Listnet, RankBoost, LambdaRank, and AdaRank). It is also worth noticing that three out of four of the best L2R algorithms are based on ensembles of learners. Compared to our baseline for this investigation, the keywords-pop recommender, most L2R algorithms improve, often significantly, with gains of up to 21.5% in terms of nDCG@10. Such an improvement demonstrates the complementarity of the various considered recommenders as features for learning effective expertise profiles.

In order to further evaluate the importance and impact of each feature on the performance of the learned models, we first sort these features by their estimated information gain [37]. Table 4 shows the results of this analysis, listing all nine considered features in increasing order of their information gain, i.e., from the least to the most informative one. From the table, we can see that the results are very consistent with the previous effectiveness results, with keyword-based recommenders having the highest information gain, which is substantially superior than that of the remaining features. In particular, the five least informative features have similarly low information gain values.

With the order imposed by the computed information gain of each feature in Table 4, we further assess their impact on the expertise profiling model learned using MART, the best performing L2R algorithm in Figure 12. In particular, starting with the MART model that uses all features, we remove one feature at a time, from the least to the most informative. At each removal step, a new MART model is learned based upon the remaining features through a 5-fold cross validation, until we have only a single feature, namely, keywords-pop. Table 5 shows the results of this experiment, in terms of nDCG@10 averaged over the test folds, along with 95% confidence intervals for the means.

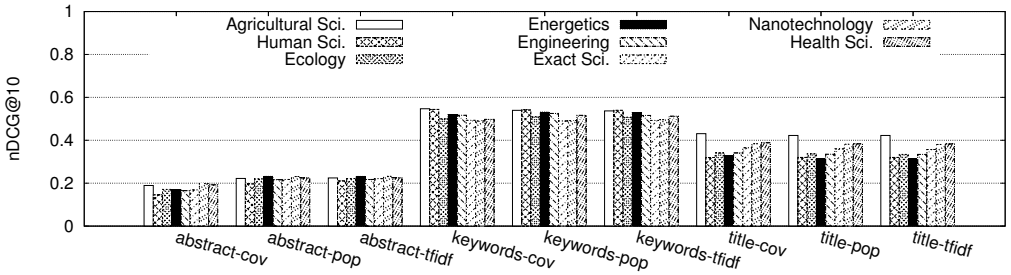


Figure 11: Profiling effectiveness for various areas of knowledge.

feature	inf. gain
title-tfidf	0.007200
title-pop	0.008010
title-cov	0.008210
abstract-tfidf	0.008900
abstract-pop	0.009730
abstract-cov	0.023060
keywords-cov	0.045410
keywords-tfidf	0.056220
keywords-pop	0.057830

Table 4: Information gain of individual features.

feature	nDCG@10
all features	0.621862 ± 0.013161
-titles-tfidf	0.621292 ± 0.013148
-title-pop	0.619677 ± 0.013174
-title-cov	0.612476 ± 0.013140
-abstract-tfidf	0.611220 ± 0.013126
-abstract-pop	0.608421 ± 0.013274
-abstract-cov	0.608024 ± 0.013301
-keywords-cov	0.605045 ± 0.013290
-keywords-tfidf	0.511833 ± 0.012279

Table 5: Profiling effectiveness after removing features.

As observed from Table 5, all features have a positive impact on the final performance of the learned model, with the least informative ones causing smaller losses when removed. This is consistent with our previous analyses. However, it is interesting to note that if we consider only the two best features for L2R, namely, keywords-pop and keywords-tfidf, the gains over the baseline are already considerable, around 18.2%, which corresponds to about 85% of the total gains obtained when using all nine features together. This result has positive impacts in terms of the trade-off "cost versus effectiveness" in the proposed L2R combination.

The results in this section demonstrate the complementarity of the aforementioned tag recommenders for identifying relevant topics of expertise. Turning back to research question Q4, these results show that an even superior expertise profiling effectiveness can be attained by leveraging these recommenders as features for a L2R algorithm. Moreover, as demonstrated through a feature removal analysis, combining a very reduced set of these features can still lead to significant gains in effectiveness, while potentially reducing the computational cost incurred by this solution.

5. CONCLUSIONS

We have investigated the suitability of tag recommendation for automatic expertise profiling, with a case study in the scientific domain. Towards this goal, we performed a large-scale user study with academic experts from all areas of knowledge working in Brazil in order to produce a comprehensive test collection for expertise profiling. With the produced test collection, we assessed the effectiveness, completeness, and robustness of the expertise profiles produced by nine different recommenders, derived from representative content-based tag recommendation algorithms from the literature applied to three different sources of expertise evidence: publication's titles, abstracts, and keywords. Our experiments demonstrate that most recommenders do satisfy such properties to a large extent, with the best ones being those based on TF or TFIDF when applied to keywords. Differently from most previous works, our analysis was conducted with a large number of real experts who can be considered as ideal evaluators for the task at hand.

The low intersection of the suggestions produced by the nine considered recommenders also motivated us to combine them by using multiple L2R strategies. To this end, we tested nine state-of-the-art L2R algorithms from the literature, with the best ones producing gains of more than 20% in effectiveness when compared to the best recommender used in isolation. Finally, we also performed an analysis of the available evidence for the rankers and found that a combination of just a couple of the best features produces most of the observed gains. As a result, the tag recommendation solutions investigated here have a potential application for real expertise profiling deployments, such as enterprises as well as academic or business-oriented social networking services.

As future work, we want to perform a historical analysis of the researchers' expertise profiles over time, using our rich dataset to better understand the evolution of topics of interest in different areas of knowledge. We also intend to further improve the effectiveness of the produced tag recommendations by using other sources of expertise evidence (e.g., coauthorship and citation networks) as well as more advanced noise filtering strategies to remove malformed tags. Finally, we plan to investigate the suitability of tag recommendation for the companion task of expert finding.

6. ACKNOWLEDGMENTS

This work is partially funded by the Brazilian National Institute of Science and Technology for the Web (grant MCT-CNPq 573871/2008-6), and by the authors' individual grants and scholarships from CAPES, CNPq, and FAPEMIG.

7. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. 2 edition, 2011.

[2] K. Balog and M. de Rijke. Determining Expert Profiles (With an Application to Expert Finding). In *Proc. of JCDL*, pages 2657–2662, 2007.

[3] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *FTIR*, 6(2–3):127–256, 2012.

[4] F. Belém, E. Martins, T. Pontes, J. Almeida, and M. Gonçalves. Associative Tag Recommendation Exploiting Multiple Textual Features. In *Proc. of SIGIR*, pages 1033–1041, 2011.

[5] F. Belém, R. Santos, M. Gonçalves, and J. Almeida. Topic Diversity in Tag Recommendation. In *Proc. of RecSys*, 2013.

[6] F. M. Belém, E. F. Martins, J. M. Almeida, and M. A. Gonçalves. Personalized and object-centered tag recommendation methods for Web 2.0 applications. *Inf. Process. Manage.*, 50(4):524–553, 2014.

[7] B. Bi and J. Cho. Automatically Generating Descriptions for Resources by Tag Modeling. In *Proc. of CIKM*, 2013.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to Rank Using Gradient Descent. In *Proc. of the ICML*, pages 89–96. ACM, 2005.

[10] S. Canuto, F. M. Belém, J. M. Almeida, and M. A. Gonçalves. A Comparative Study of Learning-to-Rank Techniques for Tag Recommendation. *JIDM*, 4(3):453–468, 2013.

[11] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to Rank: from Pairwise Approach to Listwise Approach. In *Proc. of ICML*, pages 129–136. ACM, 2007.

[12] M. de Rijke, K. Balog, T. Bogers, and A. van den Bosch. On the Evaluation of Entity Profiles. In *Proc. of CLEF*, pages 94–99, 2010.

[13] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, 2011.

[14] Y. Fang and A. Godavarthy. Modeling the Dynamics of Personal Expertise. In *Proc. of SIGIR*, pages 1107–1110, 2014.

[15] F. Figueiredo, F. Belém, H. Pinto, J. Almeida, M. Gonçalves, D. Fernandes, E. Moura, and M. Cristo. Evidence of Quality of Textual Features on the Web 2.0. In *Proc. of CIKM*, pages 909–918, 2009.

[16] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

[17] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *AOS*, pages 1189–1232, 2001.

[18] N. Garg and I. Weber. Personalized, Interactive Tag Recommendation for Flickr. In *Proc. of RecSys*, pages 67–74, 2008.

[19] P. Heymann, D. Ramage, and H. Garcia-Molina. Social Tag Prediction. In *Proc. of SIGIR*, pages 531–538, 2008.

[20] J. Lane. Let's make science metrics more scientific. *Nature*, 464(7288):488–489, 2010.

[21] Z. Lin, G. Ding, M. Hu, J. Wang, and J. Sun. Automatic Image Annotation Using Tag-Related Random Search Over Visual Neighbors. In *Proc. of CIKM*, pages 1784–1788, 2012.

[22] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. 1999.

[23] E. F. Martins, F. M. Belém, J. M. Almeida, and M. A. Gonçalves. On cold start for associative tag recoomendation. *JASIST (accepted for publication)*, 2014.

[24] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[25] R. Prokofyev, A. Boyarsky, O. Ruchayskiy, K. Aberer, G. Demartini, and P. Cudré-Mauroux. Tag Recommendation for Large-scale Ontology-based Information Systems. In *Proc. of ISWC*, 2012.

[26] C. Quoc and V. Le. Learning to rank with nonsmooth cost functions. *NIPS*, 19:193, 2007.

[27] J. Rybak, K. Balog, and K. Nørvåg. Temporal Expertise Profiling. In *Proc. of ECIR*, pages 540–546. 2014.

[28] P. Serdyukov, M. Taylor, V. Vinay, M. Richardson, and R. W. White. Automatic People Tagging for Expertise Profiling in the Enterprise. In *Proc. of ECIR*, pages 399–410, 2011.

[29] S. Siersdorfer, J. S. Pedro, and M. Sanderson. Automatic Video Tagging Using Content Redundancy. In *Proc. of SIGIR*, pages 395–402, 2009.

[30] B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. of WWW*, pages 327–336, 2008.

[31] B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. 17th International Conference on World Wide Web (WWW)*, pages 327–336, 2008.

[32] Y. Song, L. Zhang, and C. Giles. Automatic tag recommendation algorithms for social recommender systems. *TWEB*, 5:1–31, 2011.

[33] P. Venetis, G. Koutrika, and H. Garcia-Molina. On the Selection of Tags for Tag Clouds. In *Proc. of WSDM*, pages 835–844, 2011.

[34] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to Tag. In *Proc. of WWW*, pages 361–370, 2009.

[35] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[36] J. Xu and H. Li. Adarank: a Boosting Algorithm for Information Retrieval. In *Proc. of SIGIR*, pages 391–398. ACM, 2007.

[37] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of the ICML*, pages 412–420, 1997.

[38] D. Yin, S. Guo, B. Chidlovskii, B. Davison, C. Archambeau, and G. Bouchard. Connecting Comments and Tags: Improved Modeling of Social Tagging Systems. In *Proc. of WSDM*, pages 547–556, 2013.