Metadata Impact on Research Paper Similarity

Germán Hurtado Martín[1,2], Steven Schockaert[2⋆],
Chris Cornelis[2⋆], and Helga Naessens[1]

[1] Dept. of Industrial Engineering, University College Ghent, Belgium
[2] Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium

Abstract. While collaborative filtering and citation analysis have been well studied for research paper recommender systems, content-based approaches typically restrict themselves to straightforward application of the vector space model. However, various types of metadata containing potentially useful information are usually available as well. Our work explores several methods to exploit this information in combination with different similarity measures.

1 Introduction

Given the proliferation of published research results, recommending scientific papers to researchers may provide a useful complement to traditional literature search [1, 4, 5]. Various approaches may be taken to automate this task, including collaborative filtering (e.g. based on CiteULike.org or Bibsonomy.org), citation analysis (e.g. PageRank, HITS, etc.) and content-based (CB) approaches. In this paper, we focus on the latter type of systems.

Typically, CB approaches use cosine similarity applied to tf-idf vector representations of the abstracts for comparing research papers. However, various kinds of metadata are usually associated with papers, including keywords, scientific classification, journal of publication, etc.; to our knowledge, their impact on identifying related papers has not been investigated previously. In this paper, therefore, we perform an exploratory study of various methods which use such metadata directly or indirectly. Apart from assessing the relative worth of the various methods, our findings also serve to set out a baseline for future work on CB paper recommendation strategies.

2 Methodology

Test collection To build a test collection for evaluating similarity measures, we crawled a portion of the ACM library[3], consisting of all articles from 23 journals in the Artificial Intelligence domain. In addition to abstract, title, authors and journal, we also extracted the entries from the ACM classification system that

[⋆] Postdoctoral fellows of the Research Foundation – Flanders.

[3] http://portal.acm.org

were assigned to the paper, its general terms (taken from a fixed thesaurus), keywords (freely chosen by the authors) and cited papers. A description of 34658 papers was thus retrieved. Our experiments are restricted, however, to the 9594 papers for which none of the extracted fields is empty.

Similarity measures The most straightforward way to measure the similarity between two papers is by comparing their abstracts in the vector space model (method *abstract* in Table 1); each paper is represented as a vector, in which each component corresponds to a term occurring in the collection. The value of that term is calculated using the standard tf-idf approach, after removing stop words. The vectors $\mathbf{p}$ and $\mathbf{q}$ corresponding to different papers can then be compared using standard similarity measures such as the cosine, generalized Jaccard, extended Jaccard, and Dice similarity, defined respectively by

$$sim_c(\mathbf{p},\mathbf{q}) = \frac{\mathbf{p}\cdot\mathbf{q}}{\|p\|\cdot\|q\|} \qquad sim_{gj}(\mathbf{p},\mathbf{q}) = \frac{\sum_k m_k}{\sum_k M_k}$$

$$sim_{ej}(\mathbf{p},\mathbf{q}) = \frac{\sum_k m_k}{\|p\|^2 + \|q\|^2 - (\mathbf{p}\cdot\mathbf{q})} \qquad sim_d(\mathbf{p},\mathbf{q}) = \frac{2(\mathbf{p}\cdot\mathbf{q})}{\|p\|^2\cdot\|q\|^2}$$

where $\mathbf{p}\cdot\mathbf{q}$ denotes the scalar product, $\|.\|$ the Euclidean norm, $m_k = min(p_k, q_k)$, and $M_k = max(p_k, q_k)$. Alternatively, papers can be represented as vectors whose components refer to the general terms (method *g.terms*), to the keywords (method *keywords*), or to the classes of the ACM classification that have been assigned to it (method *class*). The weights are calculated analogously as in the tf-idf model. To cope with the tree structure of the ACM classification, in the *class* method, we do not only add a component for the classes at the lowest level, but also for each of their ancestors; tf-idf weighting then ensures that more emphasis is put on the lower level classes.

In the previous methods, metadata such as classes or keywords is used directly, in such a way that the most important information, the abstract, is completely ignored. We therefore follow an alternative scheme, which we refer to as explicit semantic analysis (ESA) since it is analogous to the approach from [2]. Let $\mathbf{p}$ be the vector representation obtained by method *abstract*. We now define a new vector representation $\mathbf{p_E}$ of this paper, with one component for every keyword k appearing in the collection. To define the weights of $\mathbf{p_E}$'s components, a new collection $\mathcal{C}_E = \{\mathbf{q_k} | k\ is\ keyword\}$ is first created, where $\mathbf{q_k}$ is a vector representation of the concatenation of the abstracts of all papers to which keyword k was assigned. The weights in vector $\mathbf{q_k}$ are the tf-idf scores calculated w.r.t. the new collection $\mathcal{C}_E$. The weight w_k in $\mathbf{p_E}$ corresponding to keyword k is then defined by

$$w_k = \mathbf{p}\cdot\mathbf{q_k}$$

This method is called *ESA-kw*. Similar methods are considered in which vector components refer to authors (*ESA-aut*) or to classes (*ESA-cl*). For efficiency, only authors are considered that appear in at least 4 papers in the *ESA-aut* method, and only keywords that appear in at least 6 papers in the *ESA-kw* method.

Evaluation metrics The ground truth for our experiments is derived from citations. In particular, we consider two papers as similar if either of them has cited the other, and not similar otherwise. To evaluate the performance of the methods, each paper **p** is compared against 13 others that were published in the same journal, 3 of which are actually considered similar. Similarity measures can then be used to rank the 13 papers, such that ideally the papers similar to **p** appear at the top of the ranking. In principle, we thus obtain one ranking per paper in the collection. However, since some papers are not sufficiently cited by papers that are also in the collection, only 3758 rankings were actually obtained. Their rankings can then be evaluated using standard information retrieval metrics; we use mean average precision (MAP) and mean reciprocal rank (MRR).

3 Results and Discussion

Table 1. Experimental results

	MAP				MRR			
	cos	*dice*	*e.jacc*	*g.jacc*	*cos*	*dice*	*e.jacc*	*g.jacc*
abstract	0.581	0.581	0.581	0.594	0.724	0.723	0.723	0.741
g.terms	0.367	0.367	0.368	0.367	0.443	0.443	0.444	0.442
keywords	0.472	0.469	0.470	0.475	0.634	0.631	0.629	0.634
class	0.432	0.430	0.429	0.420	0.545	0.543	0.538	0.528
ESA-aut	0.505	0.505	0.505	0.518	0.643	0.643	0.643	0.674
ESA-cl	0.535	0.535	0.535	0.527	0.667	0.667	0.667	0.673
ESA-kw	**0.597**	**0.597**	**0.597**	0.553	**0.748**	**0.748**	**0.748**	0.704

Table 1 summarizes the results of the experiment. A first important conclusion is that a content-based approach to finding related papers appears to be reasonable, as witnessed by the relatively high MAP and MRR scores of the best performing configurations. Another obvious conclusion is that all the other methods are worse or comparable to the traditional approach, *abstract*, although surprisingly the generalized Jaccard performs significantly better than the popular cosine method (paired t-test, $p < 0.001$). On the other hand, except for *g.terms* all of the methods perform substantially better than random (MAP 0.367, MRR 0.453). As could be expected, general terms are not sufficiently focused to help finding related papers. The keywords and ACM classification do seem to be useful, although alone they cannot beat *abstract.* Intuitively, keywords may be too specific, and the ACM classes too general to derive more accurate similarity information. It therefore seems promising to investigate methods that combine ACM class information with available keywords. Future work will also focus on improving the *keywords* and *class* methods by taking dependencies among keywords/classes into account (e.g. based on fuzzy rough sets, as proposed in [3]).

The ESA methods, in general, seem to outperform their "classical counterparts". However, these methods are computationally considerably more demanding, and while they make use of the abstract information, they do not succeed in improving *abstract* substantially. These results therefore suggest that the ideas behind ESA are not particularly suitable in this context. However, initial results indicate that the relative performance of the ESA methods strongly depends on the size of the test collection. In future work we will investigate the influence of the size of the test collection in more detail, as well as the role of the specific evaluation task. For instance, ground truth can be obtained using other methods than citation. An interesting idea is to derive it from user profiles from CiteULike as in [1].

As several types of metadata clearly show potential, it seems promising to consider methods for automatically learning to rank papers, based on a combination of abstract information and other features.

References

1. Bogers, T., Van den Bosch, A.: Recommending scientific articles using CiteULike. Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08), 287–290 (2008)
2. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 1606–1611 (2007)
3. Hurtado Martín, G., Cornelis, C., Naessens, H.: Personalizing information retrieval in CRISs with Fuzzy Sets and Rough Sets Proceedings of the 9th International Conference on Current Research Information Systems (CRIS2008), 51–59 (2008)
4. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. Proceedings of the 2002 ACM conference on Computer Supported Cooperative Work (CSCW'02), 116–125 (2002)
5. Proceedings of ECML PKDD (The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) Discovery Challenge 2009, Bled, Slovenia, September 7, 2009.