Scientometrics
https://doi.org/10.1007/s11192-019-03125-9

Discovering related scientific literature beyond semantic similarity: a new co-citation approach

Oscar Rodriguez-Prieto[1] · **Lourdes Araujo**[2,3] · **Juan Martinez-Romo**[2,3]

Received: 11 September 2018

Abstract
We propose a new approach to recommend scientific literature, a domain in which the efficient organization and search of information is crucial. The proposed system relies on the hypothesis that two scientific articles are semantically related if they are co-cited more frequently than they would be by pure chance. This relationship can be quantified by the probability of co-citation, obtained from a null model that statistically defines what we consider pure chance. Looking for article pairs that minimize this probability, the system is able to recommend a ranking of articles in response to a given article. This system is included in the co-occurrence paradigm of the field. More specifically, it is based on co-cites so it can produce recommendations more focused on relatedness than on similarity. Evaluation has been performed on the ACL Anthology collection and on the DBLP dataset, and a new corpus has been compiled to evaluate the capacity of the proposal to find relationships beyond similarity. Results show that the system is able to provide, not only articles similar to the submitted one, but also articles presenting other kind of relations, thus providing diversity, i.e. connections to new topics.

Keywords Scientific related literature · Recommendations · Co-citation · Statistical model · Semantic similarity

✉ Lourdes Araujo
lurdes@lsi.uned.es

Oscar Rodriguez-Prieto
rodriguezoscar@uniovi.es

Juan Martinez-Romo
juaner@lsi.uned.es

[1] Computational Reflection Research-Group, Universidad de Oviedo, Oviedo, Spain

[2] Natural Language Processing and Information Retrieval Group, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal, 16, 28040 Madrid, Spain

[3] IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, 28019 Madrid, Spain

Published online: 17 May 2019

Introduction

Some years ago we tackled the problem of broken links in web pages. We looked for methods to automatically recover those links. As usual, the first step in our research was searching for related literature. We found very few works devoted to the problem and some related aspects. Only after several months, and by chance, we found that there were some other related works, but they refer to the problem as "missing pages".

During the course of another study on spam filtering we discovered the recent deep learning models for machine learning. Documents related to these two fields are not highly similar in their contents, but they exhibit a certain relationship as the new learning approach is becoming more and more frequently applied to some natural language problems.

This kind of cases show how helpful can be a model that suggests related literature, using criteria beyond the similarity. The problem we consider in this work is how to capture those cases.

Several authors (Pedersen et al. 2007; Resnik 1999) have considered the distinction between semantic similarity and semantic relatedness. Thus Pedersen et al. (2007) emphasize that semantic relatedness and semantic similarity are different concepts. Semantic relatedness tries to capture the human capacity to assess the degree of relatedness between concepts. Thus an *apple* and a *tree* are highly related, even if they are not similar. Similarity is a special case of relatedness that focuses on the likeness (in the shape or form) of the concepts. A measure of semantic similarity takes as input two concepts, and returns a numeric score that quantifies how much they are alike.

Harispe et al. (2015) explore the distinction between semantic similarity and semantic relatedness, and based on the observation of previous works dealing with the subjects, they propose a definition for these concepts. On the one hand, semantic relatedness is defined as *the strength of the semantic interactions between two elements with no restrictions on the types of the semantic links considered.* On the other hand, semantic similarity is defined as a *subset of the notion of semantic relatedness only considering taxonomic relationships in the evaluation of the semantic interaction between two elements*, i.e. semantic similarity measures compare elements regarding the constitutive properties they share and those which are specific to them.

The problem we consider in this work is how to capture the cases of semantic relatedness between scientific literature that are not captured by measures of semantic similarity. To this purpose, in this work we propose a new co-citation based approach. Co-citation is a semantic relatedness measure for documents based on the frequency with which two documents are cited together by other documents.

However, it is not the first time co-citation has been used to recommend literature. Small (1973), one of the first authors to pay attention to the potential of this approach, said *co-citation is a relationship which is established by the citing authors. In measuring co-citation strength, we measure the degree of relationship or association between two articles as perceived by the population of citing authors*. From this point of view, co-citation can be considered a collaborative approach to recommend literature. We also think that co-citation provides a way to identify relations that can not be captured by semantic similarity. This fact makes the articles selected by the system much more diverse than other approaches.

Novelty and diversity play a central role in recommendation systems (Castells et al. 2011). In the field of recommender systems, diversity and novelty are related concepts, but they are different. Novelty refers to the degree of difference of a piece of information with

respect to what has been previously given. Diversity usually refers to the differences within a set of items. According to Castells et al. (2011), it is related to novelty since when a set is diverse, each item can be considered "novel" with respect to the rest of the set. Besides, including a novel item in a set tends to increase the global diversity. In a generic recommendation approach, the diversity in a set of items can be measured in terms of the variety, or pairwise dissimilarity, of items in the list. Novelty can be defined in terms of how many users are familiar with the items.

Using co-cites helps to provide diversity and to handle cases as those mentioned above. In the first example, two papers dealing with a same problem but using different terminology, you can expect to find other works citing both when referring to the common problem. In the second case, subjects are related by other reasons, such as a methodology to deal with a problem, two different problems tackled with the same methodology, evaluated on the same data set, etc. We can also expect to find them co-cited in works either dealing with the methodology, the problem, the collection, etc.

Though it is not the first work using co-citation to recommend literature, there is a key difference. Whereas most authors (Pohl et al. 2007; Ding et al. 2009; Mustafee et al. 2010) considered just the frequency of co-cites to determine the relevance of the relation and establishing links between cites of authors in a graph[1], we propose a new approach to measure the strength of co-citation. This makes a great difference. Highly cited articles are more likely to appear together than average [an example of the 'friendship paradox' in networks Feld (1991)]. Thus, it is essential to distinguish those cases from other in which co-citation has a significant meaning. We propose a model to measure the probability of two scientific articles are co-cited more frequently than they would be by pure chance. This relationship can be quantified by the probability of co-citation, obtained from a null model that statistically defines what we consider pure chance. Specifically, to assign a significance to the co-occurrence of two cites we use a null model according to which cites are randomly and independently distributed among the documents of the collection. Looking for article pairs that minimize the pure chance probability, the system is able to recommend a ranking of articles in response to a given article.

We have evaluated our proposal on two datasets, the ACL Anthology collection, specifically using the current release of ACL ARC (Bird et al. 2008) along with the ACL Anthology Network (Radev et al. 2013), and the DBLP dataset (Tang et al. 2008). We have also compiled a corpus, ACL_rel100, manually annotated, to evaluate the ability of our proposal to find relations beyond similarity.

The main contributions of this paper are (1) an algorithm able to capture relationships different from the pure similarity of content, (2) and able to select the co-occurrence cases that are really significant, and (3) a corpus of pairs of articles manually annotated according to their degree of relationship, which has allowed us to test the ability of our proposal to discover articles related beyond the content similarity.

The remaining of the paper proceeds as follows: Sect. 2 presents the related work; Sect. 3 is devoted to the model description; Sect. 4 describes the evaluation framework; Sect. 5 provides the results for different measures on the both corpora considered; Sect. 6 presents the ACL_rel100 corpus and the analysis of relations beyond similarity; Sect. 7 compares the proposed to others based in co-cites; finally, Sect. 8 draws the main conclusions and future work.

[1] These proposals use the relationships to contruct a graph and then they apply algorithms for graphs, such as clustering or page rank.

Related work

There are a large number of papers devoted to research-paper recommendation approaches. An in-depth review of many of the works using different approaches can be found in (Beel et al. 2016). They found that more than half of the recommendation approaches applied content-based filtering (55%). Collaborative filtering was applied by only 18% of the reviewed approaches, co-occurrence based recommender by 10%, and graph-based recommendations by 16%. Other recommendation approaches are stereotyping, item-centric recommendations, and hybrid recommendations.

Content-based recommenders (Lops et al. 2011) or content-based filtering (CBF), provide recommendations by comparing a representation of the papers contents to the representation of the user interest. These approaches usually suffer the lack of diversity problem.

The idea of Collaborative Filtering (CF) is that users like items that other users like, though this idea can be implemented in very different ways. In 1992, Goldberg et al. (1992) used the term "collaborative filtering" in a work about an experimental mail system, Tapestry, developed at the Xerox Palo Alto Research Center. They proposed collaborative filtering as a mean for people collaborating to help one another to perform filtering by recording their reactions to documents they read. Two years later, Resnick et al. (1994) used the term in a more common sense. According to them, "collaborative filters help people make choices based on the opinions of other people." They proposed GroupLens, a system for collaborative filtering of netnews, to help people find news articles. CF presents some advantages, as it does not requires the content processing of CBF, and can take advantage of the ratings provided by the users.

Recently, some proposals based on bibliometric measures have appeared. Tejeda-Lorente et al. (2015) propose to quantify the quality of both items and users without the interaction of experts, by using some bibliometric measures. Their system takes into account the measured quality as the main factor for the re-ranking of the top-N recommendations list in order to choose the latest and the best papers in a particular research area. Specifically, they use the Journal Citation Report (JCR) provided by Thomson Reuters to evaluate the quality of research resources and the h-index to evaluate the quality of researchers.

Another approach frequently used is based on the use of graphs to represent different kind of connections that exist in the scientific world. The graphs sometimes represent how papers are connected through citations (Baez et al. 2011; He et al. 2010; Liang et al. 2011). In other cases the graphs reflect the connections between authors (Arnold and Cohen 2009). Sometimes, the graphs are built considering several aspects (Zhou et al. 2008; Lao and Cohen 2010), such as cites and authors. Once the graph has been built, different graph metrics have been used to find the recommendation for a given input paper, such as random walks (Lao and Cohen 2010).

Our proposal can be included in the co-occurrence approach. Co-occurrence recommendations do not necessarily imply similarity. Items can be related by other reasons, and thus this approach is expected to provide more diverse recommendations. Small (1973) proposed the use of clusters of co-cited papers as a way to study structure of specialties in science. This work relates the strength of the relation between two co-cited papers to the raw frequency of this co-citation. Gipp and Beel (2009) use proximity of co-citations to calculate document relatedness, considering that documents cited in proximity to each other can be more strongly related. White and McCain (1998) performed an author co-citation analysis (ACA) in the Information Science discipline. The raw data are counts of the times that author pairs are cited together in articles, regardless of which of their works are cited.

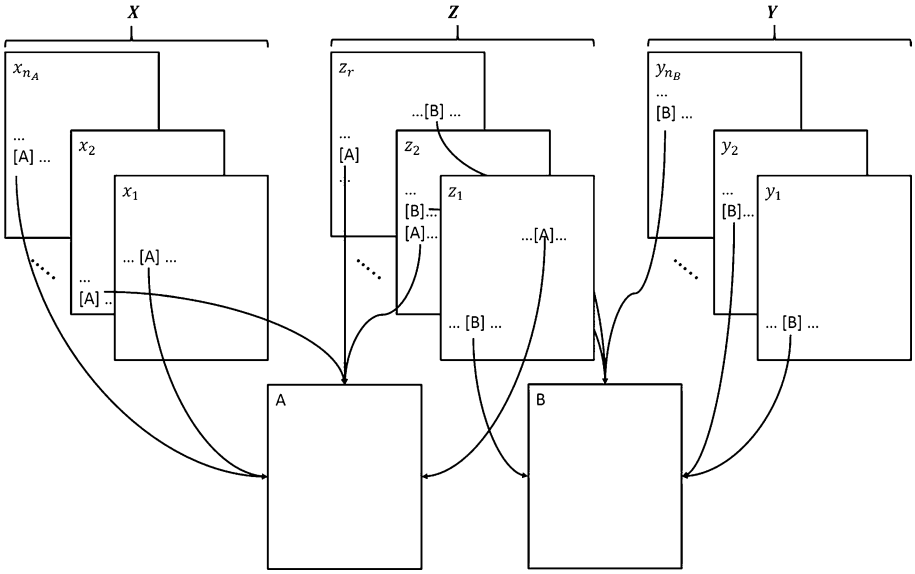

Fig. 1 Possible relations between two papers A and B. Set X corresponds to papers citing paper A but not B, set Y corresponds to papers citing paper B but not A, and set Z are papers citing both, A and B

Kim et al. (2016) also perform author co-citation analysis taking into account both citation contents and proximity. They propose a method that combines citation content with citation location to identify subject disciplines on authors. Eto (2016) proposes applying a kind of rough co-citation in order to expand co-citation networks. A rough co-citation relationship is a relation between a pair of documents that are cited by two other documents in a similar citation context.

The co-citation approach can provide diversity to the recommendation, what is so important for the user satisfaction (Kotkov et al. 2016). However, this approach requires a high degree of confidence in the recommendation to avoid selecting recommendations unrelated to the user interest. Our proposal differs from the above described ones in that we apply a model able to distinguish co-citations produced by chance from clearly related cases. It is done by computing the probability of co-citation, obtained from a null model that statistically defines what we consider pure chance.

Proposed approach

Figure 1 shows a scheme of the model. Intuitively, two particular papers, A and B, will be considered unrelated whenever the number of times they are cited independently overwhelms the number of times they are co-cited. The upper part of the figure shows three sets of papers. Set X corresponds to papers citing paper A but not B, set Y corresponds to papers citing paper B but not A, and set Z are papers citing both, A and B. The size of the set X is n_A, the index of the last paper in the set. The size of the set Y is n_B, and the size of Z is r.

Let us consider a real case corresponding to papers from the ACL collection used for evaluation. Let be paper A the one with identifier P08-1033 in the corpus, and let be B the one with identifier P07-1125. In this case n_A (papers citing only A) is 2 and n_B (papers citing only B) is

17, being r, the number of times that A and B are co-cited, 20. Thus, the number of co-cites is important compared to the number of independent cites to A and B.

A different case happens for the papers P02-1040 (A) and J93-2004 (B), two of the most cited papers in the collection. In this case n_A is 867, n_B is 904 and r is 24. Although the number of co-cited in this second case, 24, is larger than in the first case, 20, it is small respect to the number of cites of A and B, and thus we do not expect a relation between A and B. Due to the large number of cites of A and B, we can expect a high probability of being co-cited just by chance.

Let us now see how to formalize this idea by defining a null model to distinguish the statistical significant cases of co-citations.

We need to assign a significance to the co-occurrence of two cites in a certain number of documents out of the whole collection. This is akin to statistical hypothesis testing, the hypothesis being that the two cites co-occur because of semantic relatedness. Statistical hypothesis testing relies on the setting of a null model that defines what we consider pure chance. In our null model cites are randomly and independently distributed among the documents of the collection. Co-occurrence will be considered statistically significant if it is unlikely that it arises by pure chance—i.e., generated by the null model. If two cites are found respectively in n_1 and n_2 documents out of the N that form the corpus, to count in how many arrangements of two cites coincide in exactly k documents we must realize that there are four kinds of documents: k documents containing both cites, $n_1 - k$ documents containing only the first cite, $n_2 - k$ documents containing only the second cite, and $N - n_1 - n_2 + k$ documents (provided this number is non zero) containing none of the cites. Thus the sought number of arrangements will be given by the multinomial coefficient

$$\binom{N}{k, n_1 - k, n_2 - k} \tag{1}$$

Hence, the probability that two cites that appear in n_1 and n_2 documents each, and are randomly and independently distributed among N documents coincide in exactly k of them is obtained as

$$p(k) = \binom{N}{n_1}^{-1} \binom{N}{n_2}^{-1} \binom{N}{k, n_1 - k, n_2 - k} \tag{2}$$

if $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ and is zero otherwise.

We can write Eq. (2) in a more convenient form to make it computationally practical. For that purpose we introduce the notation $(a)_b \equiv a(a-1)\ldots(a-b+1)$, for any $a \geq b$, and without loss of generality assume that the first cite is the most frequent cite (i.e., $n_1 \geq n_2 \geq k$). Then

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2} (k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2 - k} (N - n_2 + k)_k (k)_k}, \end{aligned} \tag{3}$$

where in the second form we have used the identity $(a)_b = (a)_c (a-c)_{b-c}$ valid for $a \geq b \geq c$. Eq. (3) is better written as

$$p(k) = \prod_{j=0}^{n_2-k-1} \left(1 - \frac{n_1}{N-j}\right) \times \prod_{j=0}^{k-1} \frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)} \tag{4}$$

This allows us to determine a p value for co-occurrence of the two cites as

$$p = \sum_{k \geq r}^{n_2} p(k), \tag{5}$$

where r is the number of documents in the corpus where the two cites are actually found together. If $p \ll 1$ we can consider that the appearance of the two cites in the same document is significant, and therefore it is likely that their meaning is related.

This statistical model, which has been successfully applied to other Natural Language Processing (NLP) problems (Martinez-Romo et al. 2011), does not assume a normal distribution of data, as other statistical models do, and therefore is valid even when we are dealing with small numbers of cases.

Evaluation framework

For evaluation, we have used two datasets, the ACL Anthology, and the DBLP Computer Science Bibliography.

We are interested in evaluating two aspects of the articles recommended for a given article. On the one hand, we want to verify that our system actually recommends related items. On the other hand, we want to check if the system is able to capture other relations, apart from similarity. To analyse the first issue we have applied two different methodologies. We have performed a manual evaluation of the relation between a set of pairs of articles. We have also tested a number of similarity measures that allow us to identify cases in which the relation between two articles is not based on semantic similarity. To this purpose, we have studied some cases of related papers according to our system, that do not present however a high degree of similarity. The objective is to analyse whether the system has been able to detect other kind of useful relation between the articles. Later on, in Sect. 6, we specifically evaluate the ability of our system to detect relationships different from content similarity using a manually curated corpus of relationships between pairs of articles.

First of all, we have studied the appropriate threshold value for the p value used to select significant relations. For this analysis we have focused on the ACL anthology archive.

The ACL anthology archive

The ACL Anthology is a digital archive of conference and journal papers in natural language processing and computational linguistics. It serves as a reference repository of research results.

Table 1 Average and standard deviation of the p value obtained for 10 sets, each composed of 1000 pairs of articles randomly chosen from the ACL corpus

Average	SD
0.034	0.090
0.035	0.090
0.033	0.090
0.032	0.090
0.037	0.090
0.030	0.090
0.034	0.090
0.038	0.090
0.031	0.090
0.042	0.091

For evaluation we have used the February 2007 release of ACL Anthology Reference Corpus (ACL ARC) (Bird et al. 2008), that consists of:

- the source PDF files corresponding to 10,921 articles from the February 2007 snapshot of the Anthology,
- automatically extracted text for all these articles,
- 13,551 files with metadata described in the metadata/anthology-XML tree, consisting of BibTeX records derived either from the headers of each paper or from metadata taken from the Anthology website.

The metadata includes a unique ID assigned to each paper, the paper's author(s), title, publication venue, and year of publication. After deleting the repeated items, we have 20.989 items left.

As the ACL Anthology does not include any citation information, we have also used the ACL Anthology Network (AAN), a manually curated networked database of citations, collaborations, and summaries from the ACL Anthology. The ACL Anthology Network (Radev et al. 2013) was built from the original pdf files available from the ACL Anthology. AAN provides citation and collaboration networks of the articles included in the ACL Anthology. It also includes rankings of papers and authors based on their centrality statistics in the citation and collaboration networks. We have made the set used publicly available at http://nlp.uned.es/~lurdes/ACL.rar, for experiment reproducibility.

Threshold value analysis

A key parameter in our system is the threshold value for the p value to discriminate whether a relation is or it is not statistically significant.

In order to get an idea of the range of values taken by this parameter we have computed the average and standard deviation of 1000 pairs randomly chosen from the ACL corpus. Table 1 shows the results obtained for 10 different experiments. As we can see, the average of the p value obtained is around 0.03. This low value indicates that due to the nature of the corpus, it is frequent to find a certain degree of relationships between many pairs of papers. Accordingly, to be on the safe side when looking for a greater degree of relation we have to choose lower values that average. We have used 10^{-6} in most experiments.

Table 2 Human comparison of some selected relations. ID1 and ID2 are the identification tag of the considered articles. N. cites stands for the number of cites of the article of ID1 in the collection. The last column indicates the degree of relationship observed when reading and comparing each pair of articles

	ID1	ID2	*p* value	*N*. cites	Rel. degree
1	J93-2004	A00-2018	7.56E−82	928	Related
2	P02-1040	P03-1021	< 10E−257	891	Highly related
3	J93-2003	C96-2141	4.49E−208	729	Highly related
4	J03-1002	P02-1040	9.88E−222	656	Related
5	P07-2045	P02-1040	6.09E−256	591	Related
6	N03-1017	P03-1021	3.53E−277	556	Related
7	P03-1054	J98-4004	2.24E−67	394	Related
8	J96-1002	W02-2018	4.34E−21	376	Highly related
9	J86-3001	P87-1022	2.00E−38	354	Related
10	P98-2127	P90-1034	1.49E−45	305	Highly related
11	P05-1022	P08-1067	1.10E−63	290	Highly related
12	P05-1033	P03-1021	1.12E−167	283	Related
13	J02-3001	W05-0620	7.14E−98	280	Highly related
14	W02-1001	P05-1012	1.74E−67	280	Related
15	J97-3002	P01-1067	8.20E−119	268	Highly related
16	W96-0213	J93-2004	1.76E−49	267	Related
17	P97-1003	A00-2018	6.94E−79	261	Highly related
18	P98-1013	J02-3001	2.00E−88	260	Related
19	W02-1011	P02-1053	8.55E−219	254	Highly related
20	C92-2082	P99-1008	4.90E−82	248	Highly related

Manual results analysis

For this experiment we have considered the 20 most cited papers in the ACL-ARC corpus. The reason, apart from his interest for being among the most relevant, is that they are not expected to present sparsity problems. Later on, we will study the system performance on less cited papers. For each article in this set we have chosen the most related one according to our model, i.e., the one with the smallest p value. For each of these pairs, we have manually analysed the degree of relationship. This manual evaluation takes into account the whole paper. Table 2 shows the results. We can see that the articles selected by our approach in every pair are related or highly related.

Evaluating with similarity measures

In measuring similarity automatically, we have analysed two different parameters. One is the representation of the article, for which we have used the vector space model with different weighting schemes. The other is the similarity measure between the vectors representing the articles. We have used the S-Space package (Jurgens and Stevens 2010) to compute these measures. It is a software tool for building semantic spaces.

The provided algorithms deployed different similarity measures for the vectors representing the abstract of the documents. We construct these vectors applying stop word removal and stemming, and using different weighting schemes.

We have considered the following weighting schemes:

- Term Frequency (TF), the number of occurrences of a term in the abstract of the article.
- Weighted Term Frequency (WTF), the term frequency divided by the length of the abstract.
- Term Frequency-Inverse Document Frequency (TF-IDF), which combines the term frequency of a term t in a document d, with the inverse document frequency of the term in the collection D:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log \frac{|D|}{|c \in D : t \in c|}$$

where $|D|$ is the number of documents in the corpus, and c are the documents where the term t appears.

To compute similarity, we have considered the following measures:

- *Euclidean distance based similarity*. The distance between two points in Euclidean space is computed as:

$$D_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}$$

From this distance, Euclidean similarity is computed as:

$$S_E(x, y) = \frac{1}{1 + D_E(x, y)}$$

- *Cosine similarity* S_C, a measure of similarity between two vectors that measures the cosine of the angle between them:

$$S_C = \frac{\mathbf{x} \times \mathbf{y}}{|\mathbf{x}| \times |\mathbf{y}|}$$

We do not consider the WTF weighting scheme as the results with this measure are the same as for the TF weighting scheme.

- *Pearson coefficient*, a measure of the linear correlation between two variables x and y, giving a value between +1 and − 1, where 1 corresponds to total positive correlation, 0 to no correlation, and − 1 to total negative correlation. It measures the degree of linear dependence between two variables.

$$\text{Pearson}(x, y) = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \sum_{i=1}^{n} {y_i}^2 - (\sum_{i=1}^{n} y_i)^2}}$$

As in the cosine case, we do not consider the WTF weighting scheme as the results with this measure are also the same as for the TF weighting scheme. This coefficient takes values in the range $[-1, 1]$. In order to have the values in the same range [0, 1] as the previous measures we compute the Pearson similarity S_P as $|Pearson(x, y)|$.

Table 3 Comparison of weighting schemes and similarity measures for te ACL corpus. Greatest inverse correlation appears in boldface

Measure	Mean	SD	Pearson coef.
Euclidean-TF	0.05	0.009	0.005
Euclidean-WTF	0.84	0.009	− 0.07
Euclidean-TF-IDF	0.67	0.01	0.05
Cosine-TF	0.25	0.06	− **0.187**
Cosine-TF-IDF	0.12	0.05	− 0.108
Pearson-TF	0.23	0.07	− 0.186
Pearson-TF-IDF	0.10	0.05	− 0.102

Table 4 Similarity computed using TF as weighting scheme and cosine similarity. ID1 and ID2 are the identification tag of the considered articles. The last column indicates the degree of relationship observed when reading and comparing each pair of articles

	ID1	ID2	Sim.	Rel. degree
1	J93-2004	A00-2018	0.19	Related
2	P02-1040	P03-1021	0.25	Highly related
3	J93-2003	C96-2141	0.27	Highly related
4	J03-1002	P02-1040	0.17	Related
5	P07-2045	P02-1040	0.14	Related
6	N03-1017	P03-1021	0.23	Related
7	P03-1054	J98-4004	0.21	Related
8	J96-1002	W02-2018	0.28	Highly related
9	J86-3001	P87-1022	0.41	Related
10	P98-2127	P90-1034	0.16	Highly related
11	P05-1022	P08-1067	0.33	Highly related
12	P05-1033	P03-1021	0.17	Related
13	J02-3001	W05-0620	0.32	Highly related
14	W02-1001	P05-1012	0.18	Related
15	J97-3002	P01-1067	0.24	Highly related
16	W96-0213	J93-2004	0.30	Related
17	P97-1003	A00-2018	0.31	Highly related
18	P98-1013	J02-3001	0.33	Related
19	W02-1011	P02-1053	0.19	Related
20	C92-2082	P99-1008	0.12	Related

Similarity results

For evaluating similarity we have considered again the 20 most cited papers in the ACL-ARC corpus. For each of these articles we recover the five most related articles, i.e. the 5 with the lowest probability of being related to the submitted one by pure chance.

The threshold value used to select the significant relations has been set to 10^{-6}.

Table 3 shows a summary of the results for each combination of weighting scheme and similarity measure. The first column shows the mean for the 20 considered articles of the average similarity of each article and its 5 most similar articles. The next column shows the corresponding standard deviation. The last column shows the mean for the 20 considered articles of the Pearson coefficient between the p value and the similarity value for each of the 5 pairs. We can observe a very low Pearson correlation for the Euclidean measure. The

Table 5 Comparison of similarity measures for the DBLP sample1 dataset. Greatest correlation appears in boldface

Measure	Mean	SD	Pearson coef.
Euclidean-TF	0.05	0.01	− 0.05
Euclidean-WTF	0.83	0.09	− 0.06
Cosine-TF	0.18	0.11	**− 0.21**
Pearson-TF	0.29	0.15	0.18

largest inverse correlation value is obtained using TF as weighting scheme and cosine similarity, though the value obtained using Pearson as similarity measure is very similar. The inverse correlation indicates that the lower the p value, the greater the similarity. Accordingly, we will use the combination of Cosine and TF as similarity measure in the rest of the experiments. The mean (0.25) of this measure provides a reference for considering whether a pair of articles are similar or not according to this measure. Table 4 shows the similarity values computed with this measure for the pairs of papers than have been manually evaluated in Table 2. We can observe that the pairs labelled as "highly related" have in most cases a similarity value which is similar or greater than then the average of the measure. This supports our proposal of using this measure and its average as a reference for determining whether a relationship is mostly of similarity. There is however an exception in Table 4 for the pair 10. Both papers propose a word similarity measure based on the distributional pattern of words. However, one of them (P90-1034) uses terminology related to language constructions, while the other (P98-2127) uses more terminology related to thesaurus. Thus, the similarity measure has not been able to capture the relation detected by our method.

Results for the DBLP computer science bibliography

In order to assess the generality of the proposal, we have tested it on another bibliographic dataset, the DBLP Computer Science Bibliography provided by academic search system Arnetminer (Tang et al. 2008). Specifically we have used the version v8 composed of 3,272,991 papers and 8,466,859 citation relationships. After a filtering process for eliminating conference and journal names, books and some repeated articles, the model was built for 3,156,191 articles and 7,052,386 citation relationships.

In this case, we have conducted experiments using both, a set of the most cited articles, and a set of articles with few cites, which we describe below. For each article in the set of the most cited articles, we have identified the most related article, i.e. the one with the lowest probability of being related to the considered one by pure chance.

Table 5 shows a comparison of different similarity measures for this corpus. In order to compute the similarity measures we need the abstract of the articles. The huge number of articles included in this dataset, 3,156,191, makes it very complex to download all their abstracts. Therefore, we have resorted to download a sample of 6429 abstracts over a week. We call this set, publicly available at http://nlp.uned.es/~lurdes/DBLP-sample1.rar, DBLP sample1 dataset. The second column in Table 5 shows the mean for the 20 most cited articles in the sample1 DBLP dataset of the average similarity of each article and its 5 most related articles. The next column shows the corresponding standard deviation. The last column shows the mean for the 20 considered articles of the Pearson coefficient between the p value and the similarity value. We can see that the combination TF and Cosine provides

Table 6 Some pairs of articles in DBLP Computer Science Bibliography. One of them is among the 20 most cited ones, and the other is the most related to the first one according to our proposal. The last column shows the similarity obtained using TF as weighting scheme and cosine

Pair ID	Article 1	Article 2	*p* value	Sim.
1	Distinctive Image Features from scale-invariant keypoints	PCA-SIFT: a more distinctive representation for local image Descriptors	8.24E−856	0.16
2	Fast algorithms for mining association rules in large Databases	Scalable algorithms for association mining	9.52E−510	0.31
3	The anatomy of a large-scale hypertextual web search engine	Automatic resource compilation by analyzing hyperlink structure and associated text	3.15E−604	0.07
4	Latent dirichlet allocation	The author-topic model for authors and documents	3.52E−603	0.45
5	MapReduce: simplified data processing on large clusters	Pregel: a system for large-scale graph processing	3.23E−782	0.19

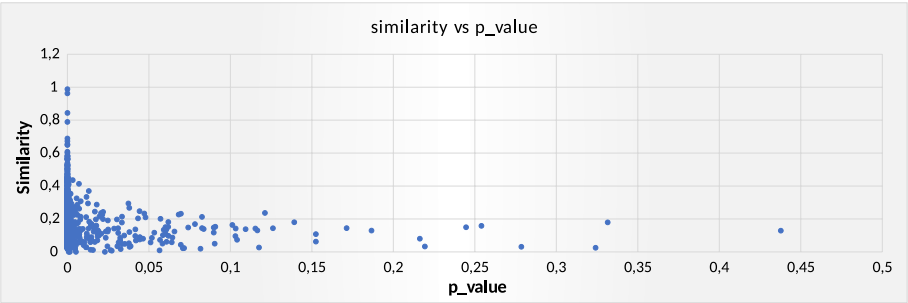

Fig. 2 Correspondence between similarity and p values for articles with few cites and co-cites. Data correspond to DBLP sample2 dataset

the highest value of inverse Pearson correlation also for this corpus. The average value of similarity for this corpus and the TF-Cosine combination is 0.18, a bit lower that the one of the ACL corpus, as we can expect because of the larger variety of topics in this corpus.

Table 6 shows the five most cited articles in the whole DBLP along with the most related article according to our system and the p_value between both papers. We can observe that in all cases the p values are extremely low. This is a very large dataset and the relations found tend to be even more significant than in the ACL Anthology. The last column in the table shows the similarity value obtained for each pair using TF and Cosine. We can observe that the similarity obtained for pairs 2 and 4 is quite high respect to the average 0.18, but it has an extremely low value for pair 3, and a value below the average for pair 1. However, all the pairs obtained are highly related. For example, in the first pair in Table 6 both papers are devoted to extract highly distinctive features from images for the image recognition process. In the second pair both papers present different algorithms for mining association rules. The first paper in the third pair proposed the Google web search engine and the PageRank algorithm, while the second paper in the pair describes a different algorithm, ARC (automatic resource compiler) for automatically compiling a list of authoritative web resources on a topic. It is a case of an interesting relationship discovered by our algorithm, that has not been captured by the similarity measure. As for the fourth pair, both papers concern the latent Dirichlet allocation (LDA) statistical model for discovering the topics involved in a document. Actually, the second paper in the pair proposes an extension of the model presented in the first one. Finally, the fifth pair is composed of two papers presenting different models for fast processing of large amounts of data. The first one presents the MapReduce programming model, which allows programs written in a functional style to be automatically parallelized and executed on a large cluster of machines. The second paper in the pair proposes a model specifically designed for distributed processing of large scale graphs, which are not easily treated with models such MapReduce. Thus, this pair is another example of interesting relationship going beyond similarity. The remaining of the 20 pairs considered are also highly related.

In order to asses both, the generality and the coverage of the model we consider now articles with relatively few cites and co-cites. We have focused on the DBLP corpus because it has a greater variety of topics and consequently a greater variety in the degree of relationship between articles. For these experiments, we have selected from the DBLP corpus articles in the following way: We randomly choose 4 articles with C citations where C varies between 1 and 11. For each of them, we randomly choose four articles, each of them

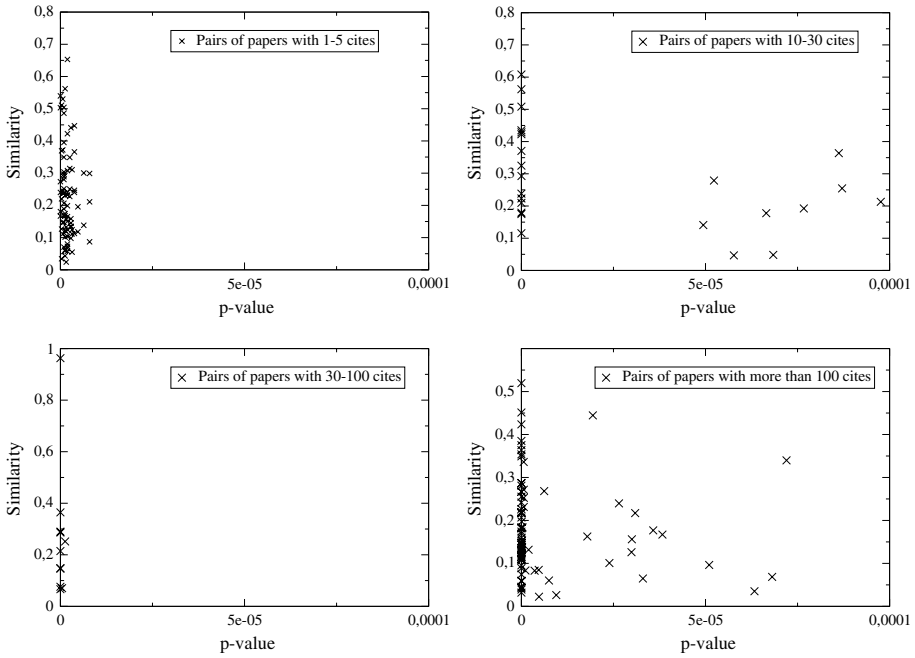

Fig. 3 Correspondence between similarity (cosine and TF) and p value for pairs with a number of citations in different ranges. Data correspond to DBLP sample2 dataset

with a relation with the first one with the p value in a different range: 0.5–0.05, 0.05–0.01, 0.01–10E−3, 10E−3 to 10E−4, 10E−4 to 10E−5, 10E−5 to 10E−6, 10E−6 to 10E−7 and finally smaller than 10E−7. Notice that it is possible that for a given range of p values we may find less than four articles that meet the condition or even none at all. With this process we have collected 670 articles, among which we have found 1116 pairs with at least a citation co-occurrence. We call this set, which is publicly available at http://nlp.uned.es/~lurdes/DBLP-sample2.rar, DBLP sample2 dataset.

Figure 2 represents the similarity between the pairs in this collection and the corresponding p value given by our model. We can observe that even including articles with low number of cites, high values of similarity correspond to low p values, thus showing the generality of the model.

These results indicate that even with relatively few cites and co-cites the model is able to capture the relations between the articles, i.e. the capacity of the method to recommend non-well cited literature. Figure 3 shows the results with more detail, separating them by ranges of number of citations (the number of citations of both papers is in the selected range). We can observe that even in the lowest range corresponding to pairs with a number of citations between 1 and 5 (and co-occurring at least once), pairs with high similarity values concentrate around low values of p value. The same happens for the other ranges of citations.

An important parameter of recommender systems (Ge et al. 2010) is coverage. According to the literature, coverage can be viewed as the percentage of the items for which the system is able to generate a recommendation. A possible measure of the prediction coverage of a system is I_p/I where I_p denotes the set of items for which a prediction can be made and I denotes the set of available items.

Table 7 Serendipity for different thresholds of p values for the DBLP sample1 dataset. The prediction model of reference has been cosine similarity above 0.18. The first column shows the threshold of p value considered for our system. The second column corresponds to the number of recommendations (RN) considered. The third and forth columns correspond to the number of recommendations provided by our system and the number of recommendations of our system that have not been provided by the similarity model, respectively. Finally the last column presents the serendipity value

p value	RN	RS	$\|RS - PM\|$	Serendipity
10E−3	1	938	869	0.92
	3	1796	1646	0.91
	5	2306	2086	0.90
10E−5	1	613	554	0.90
	3	1144	1018	0.88
	5	1379	1200	0.87
10E−10	1	270	232	0.85
	3	435	361	0.82
	5	476	374	0.78

We have computed the coverage for the ACL corpus, as the rate of articles for which there is at least one co-occurrence (12,552) with respect to the total number of articles in the corpus (20,989), obtaining a value of 59.80%. This means that the system is able to provide recommendations for almost 60% of the articles. This is a reasonable amount if we take into account that many of the papers would be recent for the time in which the corpus was collected, and thus they had not been cited yet. We have not computed the value for the DBLP corpus, since given its great variety of topics, logically most pairs have no co-occurrences between them and the value would not be representative.

Another important measure in recommender systems is serendipity. It is related to the novelty of recommendations and its ability to make surprising recommendations. A measure of serendipity (SER) is proposed in Ge et al. (2010) as:

$$SER = \frac{|RS - PM|}{|RS|}$$

where RS indicates the recommendations provided by the proposed system, and PM the recommendations provided by a primitive prediction model.

In order to compute this measure, we have resorted to the DBLP sample1 dataset of 6429 articles for which the abstracts have been downloaded. The primitive model of reference has been the ranking given by the cosine similarity provided it is above 0.18 (the average computed for this corpus). We have considered different values for the p value threshold as well as different numbers of recommendations (1, 3 and 5), taken according to the considered model ranking.

Table 7 shows the serendipity rate obtained using our system with different p values and different numbers of recommendations, for both, our model and the primitive model based on similarity. We can observe that the more restrictive the p value the lower the serendipity. For very restrictive p values, many results involve a high similarity, and thus the rate of surprising recommendations decreases. This indicates the suitability of not using excessively restrictive p values. Naturally, the greater the number of recommendations considered, the lower the serendipity value, since the reference model is capable of capturing some additional case. However, the differences are small.

Table 8 Data corresponding to the ACL_rel100 corpus of pairs of articles. The first column indicates the degree of relationship observed when reading and comparing each pair of articles: highly related (HR) or related (R). Second column shows the number of pairs found highly related (HR) and related (R). Next column shows the average p value, and the corresponding standard deviation. Last column corresponds to the average similarity computed using TF as weighting scheme and cosine similarity

	Number	Av. p value (SD)	Av. similarity (SD)
HR	82	5.13E−8 (2.76E−14)	0.088 (6.50E−4)
R	18	7.35E−8 (3.28E−14)	0.083 (8.09E−4)

Table 9 Some pairs of articles in the corpus ACL_rel100 with a relation beyond similarity. The last column shows the similarity measure given by the combination of Cosine and TF

Pair ID	Article 1	Article 2	p value	Sim.
1	Effective self-training for parsing (N06-1020)	Weakly supervised natural language learning without redundant views (N03-1023)	9.97E−07	0.13
2	What's in a translation rule? (N04-1035)	Coarse-to-fine n-best parsing and MaxEnt discriminative reranking (J03-4003)	5.76E−07	0.069
3	Recognizing contextual polarity in phrase-level sentiment analysis (H05-1044)	Accurate unlexicalized parsing (P03-1054)	5.76E−09	0.05

Beyond similarity

We will now present results that reveal the potential of the model to discover relationships beyond the similarity. Sometimes two articles showing a statistically significant relation according to our system, do not present a high degree of similarity. This means that they have been frequently co-cited—and not just because they are frequently cited individually—, but they deal with somewhat different subjects or they used different terminology to refer to the same concepts. To assess this capacity of the presented method we have constructed a corpus, ACL_rel100, composed of 100 pairs of articles from the ACL-ARC corpus. The pairs have been chosen to have a p value below 10^6, thus ensuring they are considered related by our method, and, with low similarity computed using TF as weighting scheme and cosine similarity. Specifically, after selecting those pairs with p value below 10^6, we have ranked them respect to the cosine-TF similarity, chosen those with lower similarity values. That is, they correspond to cases which will have not been predicted as related by similarity measures. The corpus has been manually annotated by three people (computer science scientists)[2], assigning to each pair one of the labels, "highly related", "related", or "low related". For each pair we have selecting the most voted label. There has not been any tie. The corpus, which is publicly available at http://nlp.uned.es/~lurdes/corpu

[2] The corpus has been annotated by the authors, that have a long experience in working with research papers.

s_ACL_rel100.txt, includes for each pair a small description of the aspects supporting the degree of relationship assigned.

Table 8 shows some data for the 100 pairs corpus. We have not found any pair "low related". We can observe that most pairs in the table are highly related (HR). The p value corresponding to the relationship between the pairs is very low—indicating a high relation—, with a very low deviation too. On the contrary, the cosine-TF similarity is very low, also with low deviation, although not so much in this case. The data show that there is a large amount of relations that the content similarity measures do not capture but that actually exist and that can be captured by methods like the one proposed herein.

Let us now examine some particular cases from the ACL_rel100 corpus in which the relationships between the pair of papers reveal some interesting connections, that are not evident from the titles of the articles. Table 9 shows some of the considered pairs. In all cases the similarity measures are below the average (0.25), as it is shown in Table 3. Of course, all of them present some degree of similarity, as they are related to NLP, but the similarity value is clearly lower than the average. However, according to our model, the corresponding p value indicates a relation between the articles of each pair.

Observing the titles of the first pair, N06-1020 and N03-1023, we can see that the first one is focused on parsing. Specifically, it proposes applying a discriminative re-ranker that reorders the list of parses produced by a generative parser. On the other hand, the second paper investigates the application of weakly supervised learning algorithms to the task of noun phrase co-reference resolution. The titles of these articles do not show that there is a relationship between them. Techniques based on the similarity of the content, as those described above, are not able to capture the relationship between them, either. However, the relation exists. In fact, both of them use self-training as semi-supervised learning technique.

Let us consider the second pair (N04-1035, P05-1022), in Table 9. The first article deals with formal semantics for word-level alignments defined over parallel corpora, the proposal relying on syntactic transformation. The second article describes a method for constructing sets of 50-best parses based on a coarse-to-fine generative parser. The method allows selecting the best parse from the set of parses for each sentence. Accordingly, though the articles deal with different applications, both of them focus on reorganizing parse tree fragments. In this way, our method has captured a relation that the content similarity approaches do not detect and which is not evident from the titles of the articles.

Finally, let us observe the last pair (H05-1044, P03-1054) in the table. The first article focuses on automatically identifying the contextual polarity for a subset of sentiment expressions. The authors apply machine learning with a variety of features, including word features. The second article proposes parsing with unlexicalized PCFGs (probabilistic context-free grammars). Thus both articles deal with very different subjects. However, the second one uses new annotations for the words which improve the parsing, and that may also be useful as features in the classifier of the first article. Accordingly, this relation may have discovered an interesting path of research.

These relationships, that had not been found with similarity-based methods, can be useful for a researcher looking for information or new ideas about alternative ways to implement self-training (first pair), reorganizing parse tree fragments (second pair) or new features for classifiers (third pair). This experiment gives an idea of the potential of the proposed method for introducing diversity.

Table 10 Comparison between the Small model (*N* co-cites) and ours (*p* value). ID1 and ID2 stand for the identification tag of the considered articles

ID1	ID2	N co-cites	ID1 cites	ID2 cites	*p* value
C92-4195	C92-1019	2	3	26	4.42E−06
W07-1502	P96-1042	2	3	26	4.42E−06
W11-1802	W10-1919	2	26	3	4.42E−06
W95-0107	P01-1069	3	161	5	4.38E−06
P12-1060	P09-1027	2	3	26	4.42E−06
N03-1017	J93-2004	17	556	928	0.96
P07-2045	J93-2004	17	591	928	0.98
J93-2004	J93-2003	21	928	729	0.98
J03-1002	J93-2004	17	656	928	0.99
P02-1040	J93-2004	24	891	928	0.99

Comparison with other systems

In order to study the differences among our model and others based on co-cites, we have evaluated on the same dataset two of the most representative models. These two models are the one proposed by Small (1973) and the one by Gipp and Beel (2009). Other models based on co-occurrence are a kind of combination of the considered ones, or they are focused on the co-occurrence of cites to authors (White and McCain 1998; Kim et al. 2016), instead to articles. The collection that we have used for the comparison is the ACL Anthology collection because it provides the whole articles (not only the abstracts), and they are needed to apply the Gipp and Beel model which takes the cites context into account.

The Small proposal relates the strength of the relation between two co-cited articles to the raw frequency of this co-cite. We have computed the number of co-cites of 371,236 pairs of articles from the ACL Anthology. Only 33 pairs have a number of co-cites above 100, and only 1082 above 20. Thus, Small model is not able to show clear relationships for most of the pairs. The Pearson coefficient between the number of co-cites and the *p* value provided by our model for all these pairs is − 0.07, i.e., no correlation is observed.

Table 10 shows some examples of differences between Small's model and ours. The top part of the table shows several pairs of articles for which Small's model does not establish a probable relationship and ours does. They are pairs that having few co-cites, can be related because despite having appeared few times, they have almost always appeared together. The bottom part of the table shows examples of the opposite case. They are pairs of articles with a relatively high number of co-cites (greater than 10), and therefore according to Small's model, probably related. However, looking at the total number of appearances of each article separately, we can think that the co-occurrences may have happened by chance and that they are not related pairs.

Figure 4 compares the number of co-occurrences associated to the Small model with the *p* value provided by our model for the 100 pairs of articles in the ACL_rel100 corpus. The pairs have been ordered by *p* value. We see that for these pairs of articles, for which a relation or a high relation has been manually verified, Small's model presents a great variability. In fact, for most pairs, the number of co-occurrences is below 20.

Gipp and Beel assume that the closer the cites are to each other the more likely it is that they are related. Accordingly, they weight the strength of the relationship by the proximity of the citations in the text. Specifically, if two cites appear in the same sentence the

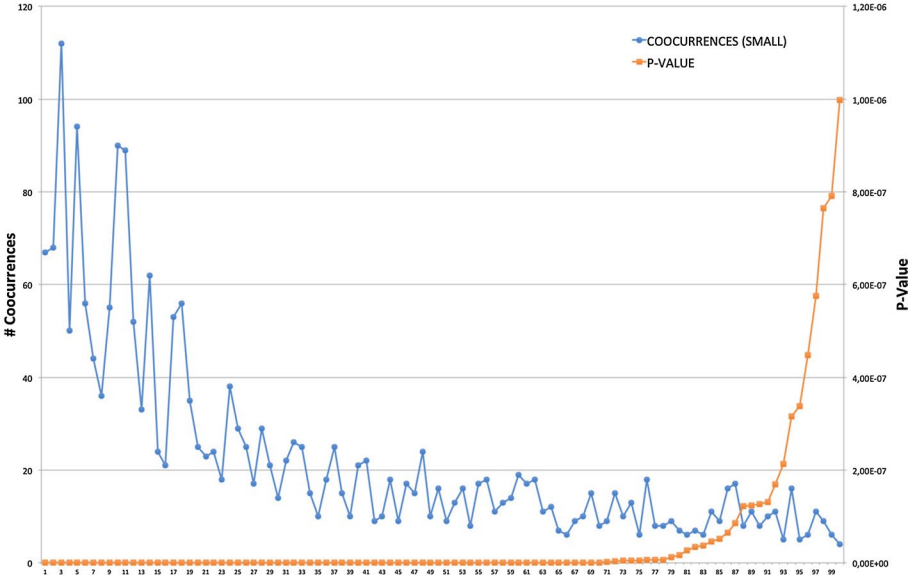

Fig. 4 Comparison of the degree of relationship provided by Small's model and ours (p value) for the pairs of articles from ACL_rel100 corpus

Table 11 Comparison between the Gipp and Beel model and ours (p value) for some pairs of articles in ACL_rel100. ID1 and ID2 stand for the identification tag of the considered articles. The number of co-cites and the number of cites of each article in ACL Anthology is also included

ID1	ID2	N co-cites	ID1 cites	ID2 cites	Gipp and Beel	p value
W06-1606	P03-2041	21	69	76	9	6.16E−36
P04-1036	P99-1004	6	79	77	1.5	4.48E−07
A88-1019	P95-1037	11	236	121	2.25	1.24E−07
P09-1063	P02-1040	10	21	891	3.25	4.15E−09
W02-0908	P99-1016	10	38	65	1.5	1.73E−17
W96-0213	A88-1019	16	267	236	3.25	6.47E−08
J08-4003	P06-1055	8	45	189	3	6.02E−09
P08-1067	J97-3002	10	85	268	3	1.31E−07
P08-1067	N06-2033	10	85	44	1	1.52E−15
N06-1020	N03-1023	4	82	20	1	9.98E−07

probability that they are very similar is higher and the weighing factor is 1, it is 1/2 for cites appearing in the same paragraph, 1/4 for cites in the same section and 1/8 for cites in the same document.

We have also implemented this model on the ACL_rel100 corpus. To do this we have recovered from the ACL Anthology collection the whole articles citing the pairs of articles included in ACL_rel100, since we needed to analyze all the text in order to compute the distance between the co-occurrences.

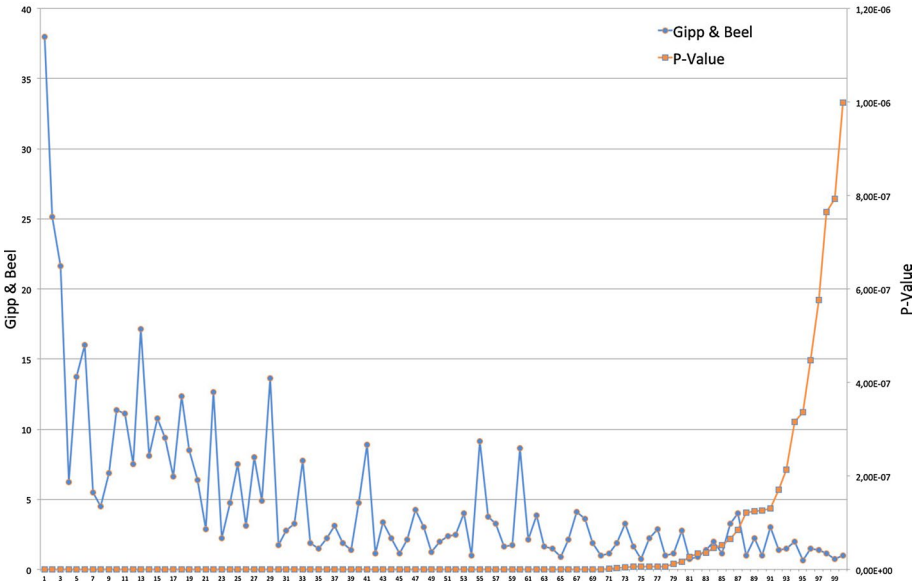

Fig. 5 Comparison of the degree of relationship provided by Gipp and Beel's model and ours (p value). for the pairs of articles from ACL_rel100 corpus

Table 11 compares the results for some of the pairs of articles in ACL_rel100. We can see that for these pairs, whose relationship has been checked manually, while our model provides p values that indicate the existence of the relationship, Gipp and Beel's model does not always capture the relationship, providing low values of relationship for some pairs such as those selected in the table.

Figure 5 compares the weighted number of co-occurrences associated to the Gipp and Beel's model and the p value provided by our model for the 100 pairs of articles in the ACL_rel100 corpus. Gipp's model smoothes the Small model's results, leading to less variability, as the figure shows. Nevertheless, the cases in which the Small model does not capture an existing relationship are not captured by the Gipp and Beel model either.

We can conclude that our model captures different relationships from those captured by the other models considered. The main difference is that our model takes into account, not the absolute co-occurrence of a pair of articles, but also the frequency of each of the articles in the pair.

Conclusions and future work

We have proposed a new model based on co-citations that is able to recommend highly related articles. The system can recommend articles similar to a given one in terms of terminology, but it is also able to capture other kind of relations. Thus, the proposed approach provides diversity to the recommendation. This important feature is achieved through the co-citation approach. There are other systems using co-citation, but we proposed an important refinement by selecting only those statically significant relations. Similarity relations can be distinguished from other kind of relations by means

of semantic similarity measures, as the ones used in this work. We have shown how the system is able to find interesting relations between articles that could not be found by similarity-based approaches. In addition, our system allows to adjust the degree of diversity by changing the p value threshold.

The system is efficient once the collection of articles and citations have been processed, which can be done in advance to submitting the queries to the system. The main limitation of the proposed method is common to all the methods based on co-cites, and is the cold start problem as these systems cannot predict for articles about which it has not yet collected sufficient information.

The proposed approach provides a weight for each relation. This weight, apart from being used to decide whether a relation is significant, can be used to define a graph of significant relations. We are planning to use this graph to identify other relations between articles, not necessarily co-cited. We will study the graph communities for a given article, as well as the relations between articles not directly connected in the graph but connected by a short path composed of links with a high weight. We are also planning to extend the study to collections in other scientific domains.

Acknowledgements This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE (IMIENS 2017).

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

Arnold, A., & Cohen, W. (2009). Information extraction as link prediction: Using curated citation networks to improve gene detection. In B. Liu, A. Bestavros, D.Z.Du, & J. Wang (Eds.), *Wireless algorithms, systems, and applications. Lecture Notes in Computer Science* (Vol. 5682, pp. 541–550). Berlin Heidelberg: Springer.

Baez, M., Mirylenka, D., & Parra, C. (2011). Understanding and supporting search for scholarly knowledge. In *7th European computer science summit, Milano, Italy* (pp. 1–8).

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338.

Bird, S., Dale, R., Dorr, B. J., Gibson, B. R., Joseph, M., Kan, M. Y., et al. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *European Language Resources Association (LREC)* (pp. 1755–1759).

Castells, P., Vargas, S., & Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International workshop on diversity in document retrieval (DDR 2011) at the 33rd European conference on information retrieval (ECIR 2011), Dublin, Ireland*. http://ir.ii.uam.es/rim3/publications/ddr11.pdf. Accessed 10 May 2019.

Ding, Y., Yan, E., Frazho, A. R., & Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *JASIST*, *60*(11), 2229–2243.

Eto, M. (2016). Rough co-citation as a measure of relationship to expand co-citation networks for scientific paper searches. *Proceedings of the Association for Information Science and Technology*, *53*, 1–4.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, *96*(6), 1464–1477.

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on recommender systems (RecSys '10)* (pp. 257–260). ACM

Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA)—A new approach for identifying related work based on co-citation analysis. In B. Larsen & J. Leta (Eds.), *Proceedings of the 12th international*

conference on scientometrics and informetrics (ISSI'09), international society for scientometrics and informetrics, Rio de Janeiro, Brazil (Vol. 2, pp 571–575). iSSN:2175-1935.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, *8*(1), 1–254.

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th international conference on world wide web (WWW'10)* (pp. 421–430). New York, NY: ACM.

Jurgens, D., & Stevens, K. (2010). The s-space package: An open source package for word space models. In *Proceedings of the ACL 2010 system demonstrations (ACLDemos '10), Association for Computational Linguistics, Stroudsburg, PA, USA* (pp. 30–35).

Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics*, *10*(4), 954–966.

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*(C), 180–192.

Lao, N., & Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, *81*(1), 53–67.

Liang, Y., Li, Q., & Qian, T. (2011). Finding relevant papers based on citation relations. In Proceedings of the 12th international conference on web-age information management (WAIM'11) (pp. 403–414). Berlin: Springer.

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 73–105). New York: Springer.

Martinez-Romo, J., Araujo, L., Borge-Holthoefer, J., Arenas, A., Capitán, J. A., & Cuesta, J. A. (2011). Disentangling categorical relationships through a graph of co-occurrences. *Physical Review E*, *84*, 046108. https://doi.org/10.1103/PhysRevE.84.046108.

Mustafee, N., Dwivedi, Y. K., Bell, D., & Williams, M. D. (2010). A methodology for profiling literature using co-citation analysis. In *Sustainable IT collaboration around the globe. 16th Americas conference on information systems (AMCIS 2010), August 12–15, 2010, Lima, Peru* (p. 359).

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, *40*(3), 288–299.

Pohl, S., Radlinski, F., & Joachims, T. (2007). Recommending related papers based on digital library access records. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries (JCDL '07)* (pp. 417–418). ACM.

Radev, D., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation,* 1–26. https://doi.org/10.1007/s10579-012-9211-2.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on computer supported cooperative work (CSCW '94)* (pp. 175–186). New York, NY: ACM.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '08)* (pp 990–998). New York, NY: ACM.

Tejeda-Lorente, A., Porcel, C., Bernabé-Moreno, J., & Herrera-Viedma, E. (2015). Refore: A recommender system for researchers based on bibliometrics. *Applied Soft Computing*, *30*, 778–791.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327–355.

Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B.L., Zha, H., & Giles, C.L. (2008). Learning multiple graphs for document recommendations. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)* (pp 141–150). New York, NY: ACM.