TOPIC DETECTION IN BIBLIOGRAPHIC DATABASES

Maria Biryukov
MINE group, University of Luxembourg, 6, Richard Coudenhove-Kalergi, L-1359, Luxembourg, Luxembourg

Keywords: Topic detection, Topic clustering, Graph properties, Knowledge discovery in data bases.

Abstract: Detection of research topics in scientific publications has attracted a lot of attention in the past few years. In this paper we introduce and compare various metrics of topic ranking, which allow to distinguish between general and focused topic terms. We use DBLP as a testbed for our experiments.

1 INTRODUCTION

Topic detection in scientific publications is an active research area in text mining and knowledge discovery in databases. Various techniques have been proposed for this purpose and range from language modeling (Wang et al., 2007; Jo et al., 2007; Diederich and Balke, 2007) to graph-based approaches and bibliometrics (Mann et al., 2006; Bird et al., 2009; Lars Backstrom et al., 2006; Mei et al., 2008). In this paper we study several metrics for ranking research topics. Our metrics are based on the topic distribution in publications and venues, and on the co-authorship relation. Using these metrics we show how to differentiate between general and specific topics. We also propose a way of grouping topics into semantically related clusters.

This paper is organized as follows: Section 2 describes the process of topic generation. In Section 3 the various ways of topic ranking are introduced. Section 4 outlines the approach for finding related topics. Section 5 presents the experiments and discusses the results. Paper's summary and synopsis of the future work are given in Section 6.

2 TOPIC GENERATION

The goal of the current paper is to extract topics and build topic clusters via the combination of three sources of information: text, co-authorship graph, and time. We start from extracting topic using publication titles which constitute the textual component for the purpose of this paper.

2.1 Extracting Topics

In this paper a *topic* is defined as a *collocation* composed of *n* consecutive words, where $2 \leq n \leq 3$. Requiring the topic components to be a collocation implies that they are semantically related, together convey a certain meaning which is different from the meaning of individual words, and the probability of their co-occurrence is higher than it would be expected if the words were independent (Manning and H.Schutze, 1999). In this context, expressions like "data mining" or "disjunctive logic programming" are examples of topics. To determine whether or not a sequence forms a collocation we apply a *likelihood ratio test for binomial distribution* (Dunning, 1993). This test belongs to the class of *hypothesis* tests where one formulates two hypotheses: *null hypothesis* which expresses the word independence, and *not-null hypothesis* under which the words are semantically related and their co-occurrence is not a chance event. The equations 1 and 2 formalize these hypotheses for the case of testing two words but can be extended for longer expressions.

$$H_0 : P(w^1 w^2) = p = P(w^2 | \neg w^1) \quad (1)$$

$$H_1 : P(w^1 w^2) = p_1 \neq p_2 = P(w^2 | \neg w^1) \quad (2)$$

By taking the ratio of the likelihoods of the two hypotheses λ one can say how much more likely one hypothesis is than the other. The null hypothesis H_0 is rejected if $p_1 \gg p_2$. It has been shown in (Dunning, 1993) that the quantity $-2log\lambda$ is asymptotically χ^2 distributed. Hence we can use the χ^2 distribution table to determine for each word sequence the confidence level of its $-2log\lambda$ value, and compare it to the

Biryukov M. (2009).
TOPIC DETECTION IN BIBLIOGRAPHIC DATABASES.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 236-242
DOI: 10.5220/0002332702360242

treshold value required for a collocation which is set tp 10.83 with confidence level $p = 0.001$. All candidates which satisfy the treshold are considered valid collocations and make up the resulting list of preliminary topics.

We discuss the topic lists in Section 5.

2.2 Topic Terms Refinement

As mentioned above we allow topic terms composed of two and three words (bi- and tri-grams further in this text). Any trigram can be seen as an extension of some bigram by one word. Presumably there are cases when $-2log\lambda$ values are sufficiently high to retain both - a bigram and its corresponding trigram(s) as topic terms. Thus we obtain terms like *"generative model"* as well as *"discriminative generative model"* and *"probabilistic generative model"*. However in some other cases selecting a trigram along with its bigrams may yield false positives. For example in *"world wide web"* only the trigram itself makes sense but neither *world wide* nor *wide web* are valid by themselves. To minimize such cases we complete the process of topic generation by applying *subsumption approach* proposed in (Sanderson and Croft, 1999) for the deriving of concept hierarchies from text. The original idea is the following: given two terms x and y, x **subsumes** y if the documents which y occurs in are a subset of the documents which x occurs in. Since x subsumes y and because it is more frequent, x is the parent of y. We adopt this idea and modify it in such a way that it serves in two different scenarios.

- **Cleaning Topic List from Meaningless Collocations.** Given a bigram x and its extension, trigram y, we **eliminate** x as having no stand alone meaning if it occurs in 80% of the documents (i.e. publication titles) which y occurs in. In other words, x is removed from the list of topics if it occurs as part of y in at least 80% of the cases. Note that we do not require a complete overlap between the occurrences of x and y. Doing so would lead to preserving a high number of meaningless bigrams just because of a few cases in which x did occur without y.
- **Defining Clusters of Lexically Related Terms.** Given a bigram x and its multiple extensions $Y = \{y_1, y_2, ..., y_n\}$, **the cluster is formed** with the central term being x, and the member terms $\{y_1, y_2, ..., y_n\}, y_i \in Y$.

After the refinement we can proceed with studying some of the topic properties.

3 RERANKING OF THE TOPICS

Since collocations are semantically meaningful units, the ranked list obtained in the way described above could already serve as a final ranked list of topics. However we consider the re-ranking due to the following observations. First of all, the two and three word collocations are generated separately, which results in two independent topic lists. Because bi- and tri-grams have different ranges of weights there is no straightforward way to compile them into one ranked list of topics without recurring to any external information. Second remark addresses the meaning of the collocation weight in general. The $-2log\lambda$ value of a topic reflects its relevance to the corpus as a whole. However it fails to capture the information about topic generality or specificity. Neither it sheds light on the topic relatedness. To overcome the lack of such information we define additional metrics for topic ranking. They are described in the following subsections.

3.1 Ranking of Topics by Citation

It is common to measure citations as an evidence of importance of an object or event. To decide on salience of a topic we define two types of citations: *citation by title* and *citation by conference*. The idea behind it is to consider every apparition of the given topic after its first occurrence as a reference (or citation) of the original topic. Note that at this point we incorporate time dimension into the analysis. To compute the new weight $weight_{t_i}$ of a topic $t_i \in T$ where T denotes the list of topics produced via the collocation extraction as described in subsection 2.1, we define:

- Citation by title $cite_{t,i}$ as a number of titles which topic t_i occurs in after the first apparition.
- Citation by conference $cite_{c,i}$ as a number of different conferences which topic t_i occurs in after its first apparition.

Then the resulting topic weight is given by the product of the two types of citations:

$$weight_{t_i} = cite_{t,i} \times cite_{c,i} \qquad (3)$$

This metric favors topics which have high counters for both, titles and conferences. Consequently we expect topics that reflect broad trends to outrank the more locally focused ones.

3.2 Ranking Topics by Co-authoship

So far only the textual and temporal informations have been used to create, refine, and re-rank the topics. The metric described in this subsection aims at distinguishing between broad and focused topics as well, but it uses co-author graph properties to do so. Intuitively more general topics will be spread among many not necessarily related to each other authors. More specific topics are expected to demonstrate an opposite behavior revealing tight co-author clusters behind themselves. The measure of co-author connectivity is captured by the *clustering coefficient* which quantifies how close the direct neighbors of a vertex are to form a complete graph (clique) (Watts and Strogatz, 1998).

To compute the topic weight in this graph-based metric we build a co-authorship graph G_t for each topic $t_i \in T$, with vertices $\{V'\}$ being the authors of all the papers which t_i occurs in, and edges $\{E'\}$ defined by the co-authorship relation between the authors in G_t. The topic weight $weight_{t_i}$ is given by the clustering coefficient of G_t, cc_{G_T}, and is computed as follows:

$$weight_{t_i} = cc_{G_T} = \frac{|E'|}{(|V'| \times (|V'|-1))/2} \tag{4}$$

where the nominator is the number of edges in G_t, and the denominator is the maximal number of edges that would have been in G_t if it was a complete clique.

We observe that such graphs are sparse: they represent a set of typically unrelated cliques. That is, the edges in G_t are mainly the ones which connect the authors of every given paper, but there are almost no edges between the authors of the different papers. However one may assume that some $v'_i, v'_j \in V'$ are connected to each other but not necessarily via particular t_i. It follows that G_t might not fully reflect the co-authorship relations between the authors related to t_i. To remedy the situation we complete the G_t with information from the global graph $G = \{V, E\}$, where $\{V\}$ are the authors of all publications listed in the bibliographical database, and there is an edge $e_{i,j} \in E$ between some v_i and $v_j \in V$ if they co-authored at least one paper. The process of building G_t is now modified in the following way: after the authors of all papers containing t_i are introduced and appropriately connected in G_t, every pair of unconnected vertices v_i, v_j is checked for having an edge in the global graph G. Should there be one, an edge $e_{i,j}$ is added to the G_t. After all the vertices $\{V'\} \in G_t$ have been checked a new clustering coefficient cc'_{G_T} is computed with the updated number of edges $\{E''\} \in G_t$. It makes sense now to check whether or not information from the graph G has changed the author connectivity in G_t. We do so by computing a new weight of t_i, $weight'_{t_i}$, which is the ratio of the two clustering coefficients, cc'_{G_T} and cc_{G_T}:

$$weight'_{t_i} = \frac{cc'_{G_T}}{cc_{G_T}} = \frac{|E''|}{|E'|} \tag{5}$$

We expect that the closer $weight'_{t_i}$ value is to 1 the more general is the term.

3.3 Ranking of Topics by $tf.idf$ Value

Term frequency - inverse document frequency ($tf.idf$) is another way of separating terms into general and specific. Introduced in (Spark, 1972) it has been widely used in the field of information retrieval. We use it here as a benchmark for the two other metrics introduced in subsections 3.1 and 3.2. The metric combines the term *salience* for the collection of documents (tf) with its *informativeness* (idf) presuming that the more focused terms will be concentrated in a fewer number of documents than more general ones which would be spread throughout the collection. We apply this metric as follows:

- term t_i = topic $t_i \in T$;
- document $d_j = c_j$, where c_j is a conference from the list of all conferences C in the database;
- $tf_{i,j}$ is the number of titles which t_i occurs in;
- cf_j is the number of different conferences which t_i occurs in.

The weight of each topic $t_i, t \in T$ is given by (6):

$$weight(i,j) = \begin{cases} (1+\log(tf_{i,j}))\log\frac{C}{cf_i} & if\ \ tf_{i,j} \geq 1 \\ 0 & if\ \ tf_{i,j} = 0 \end{cases} \tag{6}$$

where $f(tf) = (1+\log(tf_{i,j})), tf > 0$ is the dampening function. (See page 542 of (Manning and H.Schutze, 1999) for the details). We expect that more general topics will be featured not only by the high number of hosting titles but also by the high number of conferences which they occur in, as opposed to the more specific ones, grouped in relatively small number of venues.

In Section 5 we compare the results of all the three different metrics.

4 FINDING RELATED TOPICS

The question we did not deal with yet is how to identify semantically related topics. In subsection 2.2 we have briefly shown how to group them lexically. However this approach has left out semantic similarity of topics being strongly restricted to their wording. In this section we describe how we plan to extend the graph analysis suggested above to enable semantic clustering.

The underlying assumption is that authors which share some topic t_i belong to the same community and thus other topics T' they may share are possibly related to t_i. To check whether or not a topic $t_j \in T'$ is related to a topic t_i we use the updated graph G_t and the global graph G (as described in 3.2) and perform the following steps:

1. count how many authors in G_t share t_j (*internal links*, which we call $i-links$);
2. count how many authors in G share t_j (*external links*, $e-links$);
3. define $score_{t_j}$ as a ratio of $\frac{e-links}{i-links}$

The higher proportion of the internal links is the stronger t_i and t_j are related.

Note that this metric is straightforward and simple. Alternatively we can compute the strength of the topic relatedness using hypothesis testing, as a ratio of two likelihoods: $L(H_1)$ which expresses that t_i and t_j are related, and $L_(H_0)$ which says that they are not. (Likelihoods relation was used in (Jo et al., 2007) to compute the probability of a token to be a term, using citation graph built from the CiteSeer data).

Thus we may detect the following relationship: **text summarization**: {*multidocument text summarization, automatic summarization, information retrieval, text processing, ...* }, where "text summarization" is a topic in question t_i, and "multidocument text summarization", "text processing", etc., constitute a set of related topics T'.

By taking clustering coefficient into account (equation (5)) we may transform such clusters of related topics in some kind of hierarchy with more general topics being parent nodes of the more specific ones. We may combine this information with time to capture the dynamic development of a broader topic as a whole or trace the evolution of its subtopics.

5 EXPERIMENTS AND EVALUATION

In this section we discuss experiments that have been performed to test the methods described above. We focus on conference publications and use computer science bibliographic database DBLP as a test bed. Our experiments are run on the DBLP release from February 2008[1].

[1]The up-to-date versions of DBLP are available for download from http://dblp.uni-trier.de/xml/ in xml format.

5.1 Data Collection and Preparation

The xml file is parsed and the data is stored in a database. Then it is organized into two independent sets. One is intended for the collocation extraction and contains titles of conference papers. The initial list consisting of 610895 items is further preprocessed by converting to the low case, removing stop words (we use a list provided by the Lingua package (Potencier and Humphrey,)), punctuation, and titles which contain non-ASCII symbols. These constitute $\sim 2\%$ of the total number, and are mostly French and German ones with a few occurrences of the mathematical notation. The resulting list contains 599456 titles. In the second set we store complete information about the publications, including author names, title, year, and venue. It counts 610895 titles, 609053 authors, and 3996 conferences in the range of 49 years, from 1959 to 2008.

5.2 Evaluation of Topics on DBLP

The preprocessed list of titles serves as the input to the program which generates topics. (We use the NLP package for collocation extraction (Banerjee and Pedersen, 2003), with loglikehood ratio test λ as a statistic metric, and 10.83 as a cutoff weight for the $-2log\lambda$ value.) The process yields 392994 bi- and 3150332 tri-grams. Since the titles were modified during the preparation stage, not all the collocations are valid. We then conduct a post-processing which amounts to:

1. matching collocations to the original titles. Collocations that contain punctuation marks and/or stopwords, or which components fail to represent a sequence, are eliminated.
2. merging singular and plural cases into one entry;
3. subsumption, as described in subsection 2.2.

At the end of the post-processing we obtain a structure known in information retrieval as *inverted file* where for each entry the number of occurrences and an array of hosting titles are stored. The number of retained topics is reduced to 124480.

Table 1 shows some examples of the subsumption process. The first row illustrates elimination of a meaningless bigram "adaptable user". The second row is an example of a cluster which is formed around the bigram "ada programming". It is covered by the corresponding trigrams but is not eliminated. Analysis of the list of such clusters shows that many bigrams while covered by some set of trigrams have a meaning of their own and could potentially serve for topic labeling. The last row is an example of a cluster built around the bigram "application software". The

Table 1: Examples of subsumption procedure.

Bigram	Frequency	Trigram	Frequency	Covered
adaptable user	9	adaptable user interface	8	Yes
ada programming	9	ada programming environment	2	
		ada programming language	2	
		ada programming support	3	
		advanced ada programming	2	Yes
application software	39	application software development	3	
		application software systems	2	
		embedded application software	2	
		mobile application software	2	
		generic application software	2	No

topic designated by the bigram is broad enough and is not covered by the cluster members.

5.3 Experiments with Topic Re-ranking

As mentioned above the data stored in DBLP spans 49 years, from 1959 to 2008. However it can be seen from the Figure 1, that scientific activity starts to grow toward mid eighties. That is the reason why we restrict our experiments to topics which appeared no earlier than 1988. (The sharp fall of the curve toward the end of 2010 is explained by the fact that the data from 2007 – 2008 had not been completely introduced into the database by the time we downloaded the file.). Additionally we restrict the minimal topic frequency to 5 for the bi-grams, and 2 for the tri-grams.

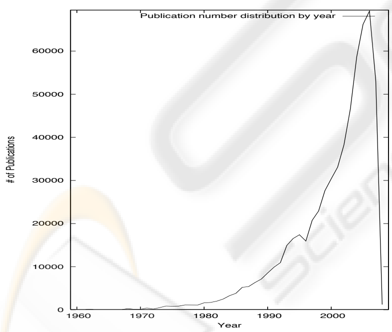

Figure 1: Paper distribution in DBLP from 1959 to 2008.

5.3.1 Results of the Ranking by Citation

Table 2 lists 20 top ranked topics according to the citation ranking computed using the equation (3).

We observe that the ranking results agree with our expectations, as almost all twenty topics designate broad areas of computer science. They are featured by high numbers of both - conferences and papers, and reflect ”trendy” research directions of the last 15years. The metric captures a high interest in relatively new topic - ”semantic web”: despite its shortest span (8 years), and relatively recent emergence (2001) it scores seventh on the total list of topics.

As we descend toward the lower ranked topics we notice that they gradually become more focused. Table 3 shows more specific topics, which may also be multi-disciplinary technical terms, like ”distance measure”. Note that the number of papers the topics occur in is still quite high while the number of conferences changes to moderate.

5.3.2 Results of the Ranking by the Clustering Coefficient

Let us now look at the topic list ranked according to the clustering coefficient cc'_{G_T} described in subsection 3.2. Table 4 shows 5 topics from the top, and 5 topics from the bottom of the list. The top ranked topics represent quite specific research fields such as theorem proving or cryptography. On the contrary **the last five topics** do not only represent the broad areas of computer science, they **correspond exactly to the top most** ranked topics according to the citation metric. This experiment proves our expectations that the clustering coefficient may serve to distinguish between broad and focused topics and gives priority to the more specific ones. We do not discuss here the ranking results yielded by the ratio of two clustering coefficients defined by equation (5). Analysis of the topic list has shown that the results do not support our predictions. Why it is so remains an open problem so far.

5.3.3 Results of the Ranking by $tf.idf$

Table 5 presents the 10 top entries from the topic list ranked according to the $tf.idf$. Since this metric gives the maximal weight to items which occur in 1 document we set the minimal number of documents (i.e. conferences in our case) to 3. We do so after the manual check of the results on an unre-

Table 2: The 20 top ranked topics by the citation metric.

topic	weight	# of conferences	# of titles	year	span
web service	2039826	654	3119	1994	13
sensor network	1777047	501	3547	1993	12
data mining	1045044	572	1827	1993	16
ad hoc network	1004598	441	2278	1995	13
wireless sensor network	648999	351	1849	1999	10
mobile agent	622362	474	1313	1994	15
wireless network	563178	371	1518	1992	17
semantic web	495624	386	1284	2001	8
multi agent system	492063	403	1221	1991	18
support vector machine	379874	341	1114	1996	13
mobile ad hoc	363025	325	1117	1998	11
virtual environment	359755	341	1055	1990	18
digital library	293112	236	1242	1991	17
association rule	261318	291	898	1993	16
face recognition	256522	251	1022	1990	18
context aware	241696	332	728	1996	12
web application	238924	322	742	1996	13
reinforcement learning	218240	248	880	1988	20
evolutionary algorithm	195487	233	839	1993	15
virtual reality	185472	288	644	1990	18

Table 3: Topics on the $500th_s$ rank.

topic	weight	# of conferences	# of titles	year	span
distance measure	6688	76	88	1990	15
heterogeneous computing	6649	61	109	1989	17
online game	6608	59	112	2001	7
aspect oriented programming	6528	64	102	1997	11
predictive control	6510	62	105	1995	11

Table 4: 5 top and 5 bottom ranked topics according to the clustering coefficient.

topic	vertices	edges (local)	edges (global)	cc'_{G_T}
spiral architecture	19	40	43	0.25146
face authentication	112	913	1030	0.16570
blue gene	209	3059	3523	0.16208
proof planning	39	53	114	0.15385
proof carrying code	21	30	32	0.15238
...				
wireless network	3311	4945	6737	0.00123
data mining	3641	5779	7563	0.00114
ad hoc network	4254	6183	8482	0.00094
web service	5732	10561	14698	0.00089
sensor network	6475	12883	16730	0.00080

Table 5: 10 top most ranked topics by the $tf.idf$.

topic	weight by $tf.idf$	# of conferences	# of papers	rank by citation	rank by clustering coefficient
research note	40.05	4	128	4289	4680
interactive presentation	34.97	4	61	7293	8121
co chair	33.92	12	135	1745	1251
output analysis	33.75	4	51	8344	2000
parallel manipulator	33.16	10	99	2581	8759
poster abstract	32.80	7	68	4536	9119
workshop chair	32.74	4	44	9229	1579
simulation optimization	32.70	7	67	4557	7423
digital government	32.16	9	76	3431	5765
low voltage	31.68	36	337	288	5568

stricted set, which put forward dozens of terms like ”session chair”, ”extended abstract”, etc. Despite this measure, we immediately notice that among the selected items there is a high number of non-topic terms such as ”research note” or ”interactive presentation”. The mixture of topic and non-topics terms happens everywhere throughout the list. Note also that the figures in the last two columns which correspond to the **topic rank** assigned by the citation and clustering coefficient metrics respectively, do not allow to establish dependency between this and the two other metrics. We explain such a behavior by the fact that $tf.idf$ is the less informed of all and clearly prefers items with the high paper-to-conference ratio which does not model the topic properties correctly.

6 SUMMARY AND FUTURE WORK

In this paper we have described the way of research topic extraction based on the titles of scientific publications. We have introduced and compared the three different methods of topic ranking aiming at distinguishing between general and specific topics. The rankings by citation and clustering coefficient have yielded topic lists which corresponded to our expectations: the first metric put forward the broader topics, while the second favored the more focused ones. On the contrary, the $tf.idf$ weighting has failed to generate a coherent list, mixing up topic and non-topic terms. Such an outcome shows that paper-to-conference relationship alone does not provide sufficient ground for the topic ranking.

So far the topic extraction is based on the publication titles only. One of the limitations of this approach is that it does not allow to capture semantic relations between the topic terms treating them as atomic. Extending textual information with the paper abstracts will alleviate this problem. It will also contribute to the process of finding related topics via the graph analysis that we have sketched in this paper.

ACKNOWLEDGEMENTS

I would like to thank Professor Christoph Schommer for his attention, valuable comments and useful suggestions while writing this paper.

REFERENCES

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381.

Bird, C., Barr, E., Nash, A., Filkov, V., Devanbu, P., and Su, Z. (2009). Structure and dynamics of research collaboration in computer science. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM.

Diederich, J. and Balke, W.-T. (2007). The semantic growbag algorithm: Automatically deriving categorization systems. In *ECDL*, pages 1–13.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, pages 61–74.

Jo, Y., Lagoze, C., and Giles, C. L. (2007). Detecting research topics via the correlation between graphs and texts. In *KDD*, pages 370–379.

Lars Backstrom, D. P. H., Kleinberg, J. M., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54.

Mann, G. S., Mimno, D. M., and McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In *JCDL*, pages 65–74.

Manning, C. and H.Schutze (1999). *Foundation of statistical natural language processing*. The MIT press, London, 2nd edition.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008). Topic modeling with network regularization. In *17 International World Wide Web Conferences (WWW)*, pages 101–110.

Potencier, F. and Humphrey, M. Lingua: Stop words for several languages. http://search.cpan.org/ creamyg/Lingua-StopWords-0.09/lib/Lingua/StopWords.pm.

Sanderson, M. and Croft, W. B. (1999). Deriving concept hierarchies from text. In *SIGIR*, pages 206–213.

Spark, J. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, pages 11–21.

Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, pages 440–442.