Detecting Source Topics by Analysing Directed Co-occurrence Graphs

Mario Kubek, Herwig Unger
Faculty of Mathematics and Computer Science
FernUniversität in Hagen
Hagen, Germany
kn.wissenschaftler@fernuni-hagen.de

Abstract: This paper describes a new method to determine the sources of topics, that influence the main topics in texts, by analysing directed co-occurrence graphs using an extended version of the HITS algorithm. Additionally, this method can be used to identify characteristic terms in texts. In order to obtain the needed directed term relations the notion of term association is introduced to cover asymmetric real-life relationships between concepts and it is described how they can be calculated by statistical means. In the experiments, it is shown that the detected source topics and the characteristic terms can be used to find similar documents and documents that mainly deal with them in large corpora like the World Wide Web. In doing so iteratively, it is possible to easily follow topics by analysing documents from these corpora using this method. This way, users can be offered this new search function in interactive search systems that goes beyond a simple presentation of similar documents. This application will be elaborated on as well.

1 Introduction and Motivation

The selection of characteristic and discriminating terms in texts through weights, often referred to as keyword extraction or terminology extraction, plays an important role in text mining and information retrieval. In [KU12] it has been pointed out, that graph-based methods for the analysis of co-occurrence graphs are well suited for keyword extraction and deliver comparable results to classic approaches like TF-IDF [SWY75] and difference analysis [HQW06]. Especially the proposed extended version of the PageRank algorithm, that takes into account the strength of the semantic term relations in these graphs, is able to return such characteristic terms and does not rely on reference corpora. In this paper, the authors extend this approach by introducing a method to not only determine these keywords, but to also determine terms in texts that can be referred to as source topics. These terms strongly influence the main topics in texts, yet are not necessarily important keywords themselves. They are helpful when it comes to applications like following topics to their roots by analysing documents that cover them primarily. This process can span several documents.

In order to automatically determine source topics of single texts, the authors present the idea to apply an extended version of the HITS algorithm [Kle98] on directed co-occurrence graphs for this purpose. This solution will not only return the most characteristic terms of texts like the extended PageRank algorithm, but also the source topics in them. Usually, co-occurrence graphs are undirected which is suitable for the flat visualisation of term relations and for applications like query expansion via spreading activation techniques.

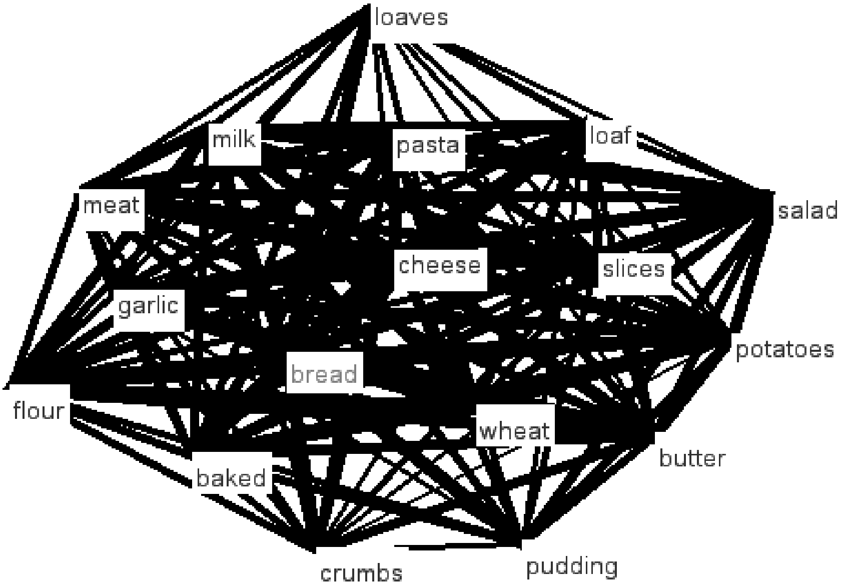


Figure 1: A co-occurrence graph for the word ”bread” (http://corpora.informatik.uni-leipzig.de/)

However, real-life associations are mostly directed, e.g. an *Audi* is a *German car* but not every *German car* is an *Audi*. The association of *Audi* with *German car* is therefore much stronger than the association of *German car* with *Audi*. Therefore, it actually makes sense to deal with directed term relations.

The HITS algorithm [Kle98], which was initially designed to evaluate the relative importance of nodes in web graphs (which are directed), returns two list of nodes: authorities and hubs. Authorities, that are also determined by the PageRank algorithm [PBMW98], are nodes that are often linked to by many other nodes. Hubs are nodes that link to many other nodes. Nodes are assigned both a score for their authority and their hub value. For undirected graphs the authority and the hub score of a node would be the same, which is naturally not the case for the web graph. Referred to the analysis of directed co-occurrence graphs with HITS, the authorities are the characteristic terms of the analysed text, whereas the hubs represent its source topics. Therefore, it is necessary to describe the construction of directed co-occurrence graphs before getting into the details of the method to determine the source topics and its applications.

Hence, the paper is organised as follows: the next section explains the methodology used. In this section it is outlined, how to calculate directed term relations from texts by using co-occurrence analysis in order to obtain directed co-occurrence graphs. Afterwards, section three presents a method that applies an extended version of the HITS algorithm that considers the strength of these directed term relations to calculate the characteristic terms and source topics in texts. Section four focuses on the conducted experiments using this method. It is also shown that the results of this method can be used to find similar and related documents in the World Wide Web. Section five concludes the paper and provides a look at options to employ this method in solutions to follow topics in large corpora like the World Wide Web.

2 Methodology

Well known measures to gain co-occurrence significance values on sentence level are for instance the mutual information measure [Büc06], the Dice [Dic45] and Jaccard [Jac01] coefficients, the poisson collocation measure [QW02] and the log-likelihood ratio [Dun93]. While these measures return the same value for the relation of a term A with another term B and vice versa, an undirected relation of both terms often does not represent real-life relationships very well as it has been pointed out in the introduction. Therefore, it is sensible to deal with directed relations of terms. To measure the directed relation of term A with term B, which can also be regarded as the strength of the association of term A with term B, the formula 1 of the conditional relative frequency can be used, whereby $|A \cap B|$ is the number of times term A and B co-occurred in the text on sentence level and $|A|$ is the number of sentences term A occurred in:

$$Assn(A \rightarrow B) = \frac{|A \cap B|}{|A|} \tag{1}$$

Often, this significance differs greatly in regards of the two directions of the relations when the difference of the involved term frequencies is high. The association of a less frequently occurring term A with a frequently occurring term B could reach a value of 1.0 when A always co-occurs with B, however B's association with A could be almost 0. This means, that B's occurrence with term A is insignificant in the analysed text. That is why it is sensible to only take into account the direction of the dominant association (the one with the higher value) to generate a directed co-occurrence graph for the further considerations. However, the dominant association should be additionally weighted. In the example above, term A's association with B is 1.0. If another term C, which more frequently appears in the text than A, also co-occurs with term B each time it appears, then its association value with B would be 1.0, too. Yet, this co-occurrence is more significant than the co-occurrence of A with B. An additional weight that influences the association value and considers this fact could be determined by

- the (normalised) number of sentences, in which both terms co-occur or
- the (normalised) frequency of the term A. The normalisation basis could be the maximum number of sentences, which any term of the text has occurred in.

The association $Assn$ of term A with term B can then be calculated using the second approach by:

$$Assn(A \rightarrow B) = \frac{|A \cap B|}{|A|} \cdot \frac{|A|}{|n_{max}|}, \text{where } 0 \leq Assn \leq 1. \tag{2}$$

Hereby, $|n_{max}|$ is the maximum number of sentences, any term has occurred it. A thus obtained relation of term A with term B with a high association strength can be interpreted as a recommendation of A for B. Relations gained by this means are more specific than undirected relations between terms because of their direction. They resemble a hyperlink on a website to another one. In this case however, it has not been manually and explicitely set and it carries an additional weight that indicates the strength of the term association. The set of all such relations obtained from a text represents a directed co-occurrence graph. The next step is now to analyse such graphs with an extended version HITS algorithm that regards these association strengths in order to find the source topics in texts. Therefore, in the next section the extension of the HITS algorithm is explained and a method that employs it for the analysis of directed co-occurrence graphs is outlined.

3 The Algorithm

With the help of the knowledge to generate directed co-occurrence graphs it is now possible to introduce a new method to analyse them in order to find source topics in the texts they represent. For this purpose the application of the HITS algorithm on these graphs is sensible due to its working method that has been outlined in the introduction. The list of hub nodes in these graphs returned by HITS contain the terms that can be regarded as the source topics of the analysed texts as they represent their inherent concepts. Their hub value indicates their influence on the most important topics and terms that can be found in the list of authorities.

For the calculation of these lists using HITS, it is also sensible to also include the strength of the associations between the terms. These values should also influence the calculation of the authority and hub values. The idea behind this approach is that a random walker is likely to follow links in co-occurrence graphs that lead to terms that can be easily associated with the current term he is visiting. Nodes, that contain terms that are linked with a low association value however should not be visited very often. This also means that nodes that lie on paths with links of high association values should be ranked highly as they can be reached easily.

Therefore, the formulas for the update rules of the HITS algorithm can be modified to include the association values $Assn$. In fact, this step is a necessity when dealing with co-occurrence graphs because otherwise less important associations would be treated like more important associations by the HITS algorithm. This is a major difference between such word nets and the World Wide Web, in which the links exist or do not exist at all.

The authority value of node x can then be determined using formula 3:

$$a(x) = \sum_{v \to x} (h(v) \cdot Assn(v \to x)) \tag{3}$$

Accordingly, the hub value of node x can be calculated using formula 4:

$$h(x) = \sum_{x \to w} (a(w) \cdot Assn(x \to w)) \tag{4}$$

The following steps are necessary to obtain a list for the authorities and hubs based on these update rules:

1. Remove stopwords and apply stemming algorithm on all terms in the text. (Optional)
2. Determine the dominant association for all co-occurrences using formula 1, apply the additional weight on it according to formula 2 and use the set of all these relations as a directed co-occurrence graph G.
3. Determine the authority value $a(x)$ and the hub value $h(x)$ iteratively for all nodes x in G using the formulas 3 and 4 until convergence is reached (the calculated values do not change significantly in two consecutive iterations) or a fixed number of iteration has been executed.
4. Order all nodes descendingly according to their authority and hub values and return these two ordered lists with the terms and their authority and hub values.

Now, the effectiveness of this method will be illustrated by experiments.

4 Experiments

4.1 Detection of Authorities and Hubs

The following tables show for two documents of the English Wikipedia the lists of the 10 terms with the highest authority and hub values. To conduct these experiments the following parameters have been used:

- removal of stopwords
- restriction to nouns
- baseform reduction
- activated phrase detection

Table 1: Terms and phrases with high authority and hub values of the Wikipedia-article ”Love”:

Term	Authority value	Term / Phrase	Hub value
love	0.54	friendship	0.19
human	0.30	intimacy	0.17
god	0.29	passion	0.14
attachment	0.26	religion	0.14
word	0.21	attraction	0.14
form	0.21	platonic love	0.13
life	0.20	interpersonal love	0.13
feel	0.18	heart	0.13
people	0.17	family	0.13
buddhism	0.14	relationship	0.12

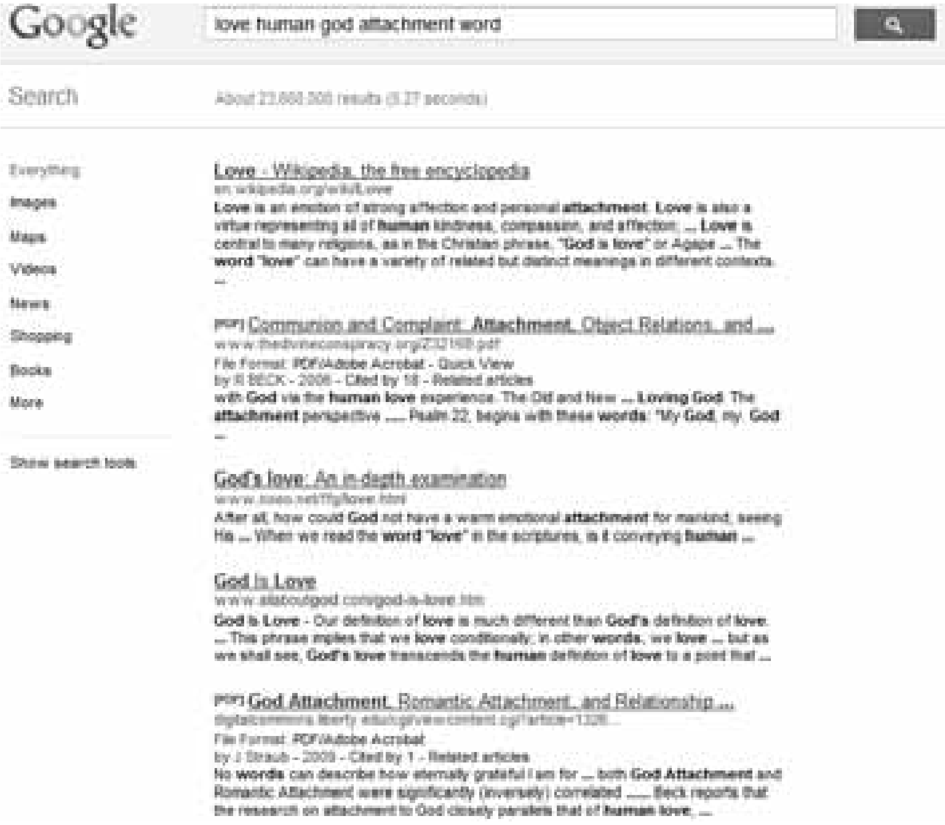
Table 2: Terms and phrases with high authority and hub values of the Wikipedia-article ”Earthquake”:

Term	Authority value	Term / Phrase	Hub value
earthquake	0.48	movement	0.18
earth	0.30	plate	0.16
fault	0.27	boundary	0.15
area	0.23	damage	0.15
boundary	0.18	zone	0.15
plate	0.16	landslide	0.14
structure	0.16	seismic activity	0.14
rupture	0.15	wave	0.13
aftershock	0.15	ground rupture	0.13
tsunami	0.14	propagation	0.12

The examples show that the extended HITS algorithm can determine the most characteristic terms (authorities) and source topics (hubs) in texts by analysing their directed co-occurrence graphs. Especially the hub list for each text provides useful information to find suitable terms that can be used as search words in queries when background information is needed to a specific topic. However, also the terms found in the authority lists can be used as search words in order to find similar documents. This will be shown in the next subsection.

4.2 Search Word Extraction

The suitability for these terms as search words will now be shown. For this purpose, the five most important authorities and the five most important hubs of the Wikipedia article ”Love” have been combined as search queries and sent to Google. Empiric experiments have shown that at most five terms and phrases should be used for this purpose. A larger number would limit the search results too much, while too few terms would return too many and possibly irrelevant results. The results of this test can be seen in figure 2 and 3.



love human god attachment word

About 23,600,000 results (0.27 seconds)

Love - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Love
Love is an emotion of strong affection and personal attachment. Love is also a virtue representing all of human kindness, compassion, and affection; ... Love is central to many religions, as in the Christian phrase, "God is love" or Agape ... The word "love" can have a variety of related but distinct meanings in different contexts. ...

[PDF] Communion and Complaint: Attachment, Object Relations, and ...
www.thedivineconspiracy.org/Z2321B.pdf
File Format: PDF/Adobe Acrobat - Quick View
by R BECK - 2006 - Cited by 18 - Related articles
with God via the human love experience. The Old and New ... Loving God. The attachment perspective ..... Psalm 22, begins with these words: "My God, my God ...

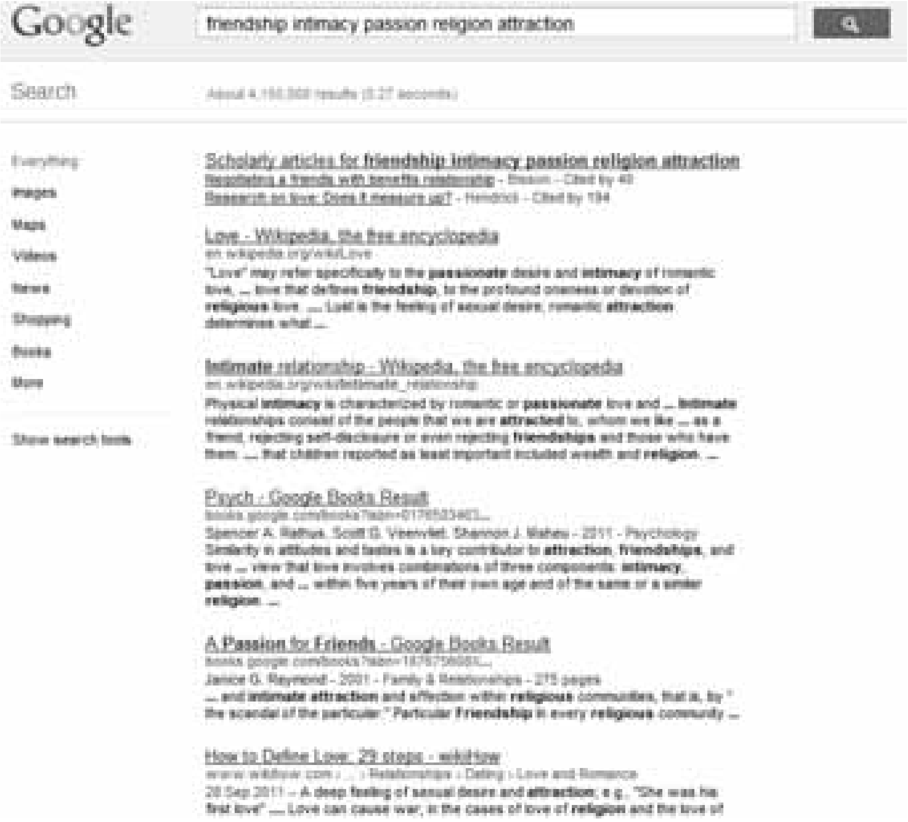
God's love: An in-depth examination
www.sass.net/?g/love.html
After all, how could God not have a warm emotional attachment for mankind, seeing His ... When we read the word "love" in the scriptures, is it conveying human ...

God Is Love
www.allaboutgod.com/god-is-love.htm
God is Love - Our definition of love is much different than God's definition of love. ... This phrase implies that we love conditionally; in other words, we love ... but as we shall see, God's love transcends the human definition of love to a point that ...

[PDF] God Attachment, Romantic Attachment, and Relationship ...
digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=1326...
File Format: PDF/Adobe Acrobat
by J Straub - 2009 - Cited by 1 - Related articles
No words can describe how eternally grateful I am for ... both God Attachment and Romantic Attachment were significantly (inversely) correlated ..... Beck reports that the research on attachment to God closely parallels that of human love. ...

Figure 2: Search results for the authorities of the Wikipedia article ”Love”

The search results clearly show, that they primarily deal with either the authorities or the hubs. More experiments confirm this correlation. Using the authorities as queries to Google it is possible to find similar documents to the analysed one in the Web. Usually, the analysed document itself is found among the first search results, which is not surprising though. However, it shows that this approach could be a new way to detect plagiarised documents. It is also interesting to point out the topic drift in the results when the hubs have been used as queries. This observation indicates that the hubs of documents can be used as a means to follow topics across several related documents with the help of Google. Hereby, it is desirable that the hubs of the analysed documents are the authorities of the found documents to obtain a chain of documents that are indeed topically depending. This possibility will be elaborated on in more detail in the next and final section of this paper.



friendship intimacy passion religion attraction

About 4,150,000 results (0.27 seconds)

Scholarly articles for **friendship intimacy passion religion attraction**
Negotiating a friends with benefits relationship - Bisson - Cited by 40
Research on love: Does it measure up? - Hendrick - Cited by 194

Love - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Love
"Love" may refer specifically to the **passionate** desire and **intimacy** of romantic love, ... love that defines **friendship**, to the profound oneness or devotion of **religious** love. ... Lust is the feeling of sexual desire; romantic **attraction** determines what ...

Intimate relationship - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Intimate_relationship
Physical **intimacy** is characterized by romantic or **passionate** love and ... **Intimate** relationships consist of the people that we are **attracted** to, whom we like ... as a friend, rejecting self-disclosure or even rejecting **friendships** and those who have them. ... that children reported as least important included wealth and **religion**. ...

Psych - Google Books Result
books.google.com/books?isbn=0176523463...
Spencer A. Rathus, Scott G. Veenvliet, Shannon J. Mahau - 2011 - Psychology
Similarity in attitudes and tastes is a key contributor to **attraction**, **friendships**, and love ... view that love involves combinations of three components: **intimacy**, **passion**, and ... within five years of their own age and of the same or a similar **religion**. ...

A **Passion** for **Friends** - Google Books Result
books.google.com/books?isbn=187875698X...
Janice G. Raymond - 2001 - Family & Relationships - 275 pages
... and **intimate attraction** and affection within **religious** communities, that is, by " the scandal of the particular." Particular **friendship** in every **religious** community ...

How to Define Love: 29 steps - wikiHow
www.wikihow.com › ... › Relationships › Dating › Love and Romance
28 Sep 2011 – A deep feeling of sexual desire and **attraction**; e.g., "She was his first love" ... Love can cause war, in the cases of love of **religion** and the love of

Figure 3: Search results for the hubs of the Wikipedia article ”Love”

5 Conclusion

In this paper, a new graph-based method to determine source topics in texts based on an extended version of the HITS algorithm has been introduced and described in detail. Its effectiveness has been shown in the experiments. Furthermore, it has been demonstrated that the characteristic terms and the source topics that this method finds in texts, can be used as search words to find similar and related documents in the World Wide Web. Especially the determined source topics can lead users to documents that primarily deal with these important aspects of their originally analysed texts. This goes beyond a simple search for similar documents as it offers a new way to search for related documents, yet it is not impossible to find similar documents when the source topics are used in queries. This functionality can be seen as a useful addition to Google Scholar (http://scholar.google.com/), which offers users the possibility to search for similar scientific articles.

Additionally, interactive search systems can employ this method to provide their users functions to follow topics across multiple documents. The iterative use of source topics as search words in found documents can provide a basis for a fine-grained analysis of topical relations that exist between the search results of two consecutive queries. Documents found in later iterations in suchlike search sessions can give users valuable background information on the content and topics of their originally analysed documents. In this context, it is also sensible to let users interactively evaluate the topical dependencies of the found documents. Highly relevant results could be marked and act as a suggestion for other users to be further examined. This function would be useful for groups or communities of like-minded people e.g. scientists in a certain field that often deal with domain-specific knowledge whose aspects have topical dependencies. A fast and correct presentation of topically depending documents would be a great help for them, especially when they have a specific information need.

Another interesting application for this method can be seen in the automatic linking of related documents in large corpora. If a document A primarily deals with the source topics of another document B, then a link from A to B can be set. This way, the herein described approach to obtain directed term associations is modified to gain the same effect on document level, namely to calculate recommendations for specific documents. These automatically determined links can be very useful in terms of positively influencing the ranking of search results, because these links represent semantic relations between documents that have been verified in contrast to manually set links e.g. on websites, which additionally can be automatically evaluated regarding their validity by using this approach. Also, these automatically determined links provide a basis to rearrange returned search results based on the semantic relations between them. These approaches will be examined in later publications in detail.

Bibliography

[Büc06] M. Büchler. Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten. Master's thesis, Leipzig University, July 2006.

[Dic45] L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, July 1945.

[Dun93] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19:61–74, March 1993.

[HQW06] G. Heyer, U. Quasthoff, and T. Wittig. *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. W3L-Verl., 2006.

[Jac01] P. Jaccard. Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[Kle98] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of ACM-SIAM Symp. Discrete Algorithms, San Francisco, California*, pages 668–677, January 1998.

[KU12] M. Kubek and H. Unger. Search Word Extraction Using Extended PageRank Calculations. In Herwig Unger, Kyandoghere Kyamaky, and Janusz Kacprzyk, editors, *Autonomous Systems: Developments and Trends*, volume 391 of *Studies in Computational Intelligence*, pages 325–337. Springer Berlin / Heidelberg, 2012.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library Technologies Project, 1998.

[QW02] U. Quasthoff and C. Wolff. The Poisson Collocation Measure and its Applications. In *Second International Workshop on Computational Approaches to Collocations*. IEEE, 2002.

[SWY75] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.