What is the value of information - search engine's point of view

Mieczysław A. Kłopotek

Institute of Computer Science of the Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa Poland

Abstract.
Within the domain of Information Retrieval, and in particular in the area of Web Search Engines, it has become obvious long time ago that there is a deep discrepancy between how the information is understood within computer science and by the man-in-the-street.
We want to make an overview of ways how the apparent gap can be closed using tools that are technologically available nowadays.
The key to a success probably lies in approximating (by means of artificial intelligence) the way people judge the value of information.

Keywords: information, semantic search.

1 Introduction

Information Science, especially Information Retrieval, and in particular Search Engines needs to measure information value for at least a couple of reasons:

- document ranking in search engines - for typical queries, consisting of a few words, hundreds or even thousands matching documents are found, so to satisfy the user, one may use additional criteria for ranking, like the value of contained information
- filtering information according to one's interests (such as fresh information, information without advertisement, formulated in a simple language, etc.),
- classification of information,
- selection of a representative of a group of documents,
- selection of information to store in a database search engine - necessary because of the enormity of the Internet as compared to the capacity of even the largest search engine databases

From user's point of view, the value of information should be measured at least along the following dimensions:

- *quality* - agreement with the state of the real world (positive for true, negative for false information)
- *utility* - usefulness for the purposes of the user (positive if useful, negative, when harmful)
- *actuality* - about the current state of the world or from the past

– *intelligibility* - whether or not the receiver can capture the content of the message
– *accessibility* - whether or not the message can reach the user for whatever reasons, (e.g. whether or not it has been sent)

Now the big question is if we are capable to capture the value of information along these dimensions. The answer is: at least partially yes. The subsequent sections will be devoted to show how with current technology these measurements may be performed.

Befolre proceeding let us however point at difficulties when matching these user expectations. Measuring information has a long tradition in Computer Science. As early as in 1948 Shannon [20] laid foundations for development of information theory quantifying the information as the entropy of the source of signal which defined the minimal bandwidth through which a given information may pass. Shannon's entropy measure was also for years considered as an upper limit for compression that only the best compression algorithms could approach. In the recent time however a new perspective is opening. In the domain of lossless compression a new branch of visually lossless compression emerged [17], bringing dozens of times higher compression rate than predicted by Shannon's information theory. With the advent of HTML and the separation of text and images also degrees of information compression far beyond original Shannon's concepts were achieved. Recently, the concept of excess entropy [7] has been coined to point at the fact that going beyond the pure statistical evaluation of the source of information (sender) by taking into account the structure (syntax), semantics, pragmatics and apobetics[1] we can achieve higher information levels than at the statistical level.

These sample developments point at two important points missed by the original proposal of Shannon: the significance of information receiver (like in the visually lossless compression) and the structure of information (like in HTML example). Hence the definition of information itself has been broadened beyond Shannon's concept and is understood as the content of a message sent from an information source (sender) to an information sink (receiver) increasing the state of knowledge of the latter. This imples that any information stems from an intelligent sender. And this is exactly the reason why the information value can be measured along the aforementioned dimensions in spote iof the fact that our computer technology is far from a through understanding of the real world - to measure information stemming from an intelligent sender we can apply artificial intelligence tools and methods as we will subsequently see.

[1] The *pragmatic level* of information means the activity of the receiver intended/achieved by the sender via the message. The issue of *apobetics* has been probably first raised by Gitt [10]. and it means the goal of the sender pursued when sending the message.

2 Measuring information quality

As mentioned, the quality of the information is the degree of compliance with the state of the real world, which a Web page describes (can be positive when the line and negative if it is inconsistent).

Can we really say how the information is consistent with the state of the real world? Not at this stage of development, but we can now at least point to situations where non-compliance is suspected :

- other users were not interested in reading it
- creator is not understanding the topic
- creator does not care about the form of the page,
- author prepared a propagandist text
- the text is in fact an advertisement

Whether or not the other users were interested in reading it, we can tell by measuring the time spent on the site and extent of returning to this page.

To detect, whether the developer worked carelessly, without much attention to content, we can measure linguistic correctness (care, the spelling, the analysis of syntax and elements of semantic analysis), as well as calculating the so-called badrank [16] to see of the author cares about the relevance of links. It may be useful to distinguish the spelling mistakes of the so-called "Typos" (e.g., by measuring the so-called edit distance properly loaded on the keyboard). Spelling errors may indicate a low level of education, and thus knowledge, and "typos" - little attention to detail paid by the creator (perhaps not read their texts). The measure should take into account not only the absolute number of errors, but the length of the text.

Web statistics can allow to distinguish among these pages with probably the highest value in terms of care and representativeness. Word documents can be assessed using methods based on tfidf (term frequency, inverse document frequency). Term frequency is the number of occurrences of words in the document, while the inverse document frequency is the inverse of the number of documents in which the term occurs (or respective logs of them). The value of the text would represent the sum of the individual words tfidf in document. Instead tfidf also the variance in the number of instances of documents and other measures of diversity of the words between documents can be used.

The above-mentioned concept of PageRank can be used to assess the value of web pages in various ways. The quality assessment is of course the basic PageRank itself. But also its variant called badrank. Badrank [16] is a measure of "spamming" by the creator as part of the so-called link farm. It is assumed that the WWWW pages pointing at the spamming page are also involved in the spamming action. Therefore, the mechanism is constructed similar to PageRank [19] except that we reverse the directions of the arrows in the Web graph. The so-called personalized random walk is assumed that is the random walker makes his out-of-boring jump to the group U of web pages which are manually qualified as spamming pages. The probability of hitting on the page is a measure of the spam. Beside spam farms [24] there are also many other techniques of spamming

and corresponding methods of spam detection, based on reverting the spamming technologies. Let us mention some, described e.g. in [18]

- inserting multiple keywords to the content - antispamming by a comparison of histogram of occurrence of words in queries and other categories of words (e.g. participation in the common words [stop-words]) in general and the distribution of popular etc. words in the document
- inserting multiple keywords to the title - antispamming as above
- inserting words that are keywords pasted together (e.g. freedownload) - antispamming by observation over the average length of words
- inserting a large part of the text as anchor text - antispamming by proper statistics
- a side effect of effort put into spamming is usually neglecting the Web page appearance - antispamming by checking for a small number of HTML tags compared to the length of text
- duplicating content - antispamming by checking for high compressibility gzip - more than four times the typical compression rate
- spam generated by a random mechanism - antispamming by computing the conditional independence of the occurrence of words in n-grams on a page

Besides spam there are also other types of harmful content on the Web, like pornography [1] (which is relatively easy to recognize), illegal trade (fake brand products), criminal, violent racist and other harassment pages, as well as advertisement pages for harmful products aiming at distorting the people and the society. In this case one relies also on a set of predefined vocabulary, focused crawling, group of editors, creating directories of harmful and / or safe pages, with support of methods of event extraction and analysis. One can assess goodness of other pages by personalized PageRank in normal Web graph and badness by personalized PageRank in the reverse graph, and then compare goodness and badness of each page.

Next one can look into such as inserted by the creator of the links are actually in line with the theme of the document. You can highlight the following techniques:

- clustering (cluster analysis), links within the cluster is considered to be topical
- Links from / to spam - to be non-topical
- matching topic subject links to the document - similarity of the text near the anchor to the vector of 20 words with the highest tfidf in the target document [5]
- as previously, but with separate treatment of links within the Web domain and going outside, classifying ranks relative to these groups (e.g., above the median)
- topical triangles: if A is related to B, A points B, and C points to A and to B, then the links from C to A and B may be considered topical.
- link between the pages A and B is topical if both pages are often found high ranking in many queries

– links of type hubs / Authorities in HITS and SALSA technology
– comparative analysis of the structure of the web pages according to the recommendations of various guidelines of web page construction (e.g. elimination of navigational links by identifying their permanent place)

In most of techniques for detecting spam or other harmful content some numerical values are computed. These values can be the ratios of frequency of occurrences as compared with the "gold standard" or results of some statistical tests of significance aiming to reject the hypothesis that the website does not represent the spam.

Beside looking for signs of page misconduct, also efforts were made to define positive tests of page quality. So in [9] it is proposed to consider a page a good one if there are no errors of any kind, if the creator of the web page can be clearly identified (with contact details), verifying that the site belongs to a serious organization having experts in the field presented by the Web page, and the inclusion of references (links) to relevant pages on the topic. Here, of course, it is necessary to use a shallow analysis of natural language, identifying named entities and the use of appropriate semantic resources (lists of individuals or organizations or types of organizations that are trustworthy). Also one needs methods for appropriate classification of the content of the page [3] to match it against the list of experts. [8] proposes a number of methods for assessing the quality of Web pages edited by communities, in which the method of time series analysis of changes and of the list of readers / writers is exploited. Unfortunately, not all suggestions of [8] are plausible, just to mention the requirement "Neutral point of view". Though today technology is close to properly assessing the attitude of the author (via called a sensual analysis) and "neutrality" can in principle be measured automatically, this measure has nothing to do with the quality measurement as our primary goal is to ensure that the contents are true (reflect actual state of the world) and hence it cannot be not neutral with respect to the truth.

3 Measuring utility

Utility is the degree to how much information brings the viewer to learn about the world to the extent necessary for the running existence / immediate action (can be positive if the beneficial and negative when it is harmful).

The difference to the quality lies in the fact that we now care for receiver's needs. So high quality information is not enough, it may prove useless for the current goals, therefore reading it may be waste of time and hence harmful.

Hence, the role of search engines is to provide the users with what they currently need. Thus, the utility can be seen in the context of the user's query.

Utility may be again judged by viewing reaction of other receivers. For example, if page advertised product and the customer bought it, it means that the information provided was useful.

Previously there were efforts to profile the user claiming that the content returned by the search engine should best fit user's profile. Nowadays such an

approach is deemed to be a failure because we cannot predict from the past user's information need in the future (the user can look now for something else).

Therefore a different pathway is steered at. Rather then trying to find user's profile, a search engine reply with diverse topics is aimed at. One presents best fitting the query but differing from the preceding ones on the list, not only by content but also by topic (in fact similarity to a query is weighted against the difference to the predecessors). So the value of a page is not only judged by its similarity to the query but also its new content compared to what is provided by better pages.

In general, measuring utility requires climbing up to at least semantic level of the message. Our group is engaged in developing a semantic search engine covering the whole Polish Internet. To ensure appropriate quality of responses, we seek a well-founded method of matching user's query to the document contents. Let us briefly explain the notion of semantic search engine and its impact on ranking method requirements.

As already stressed, semantics of information expresses the meaning of this information. In linguistics research on the semantics tries among others to relate symbols (like words, phrases, characters) to (real) beings which they mean (so-called denotations) therefore related areas like morphological and syntactic analysis is engaged. Understanding of semantics may prove useful in comprehending pragmatics (the expected acting of the recipient upon obtaining the information) and apobetics (the goal of the sender when sending the information).

Identification of the meaning of an information has been subject of intense research. So-called "semantic search" is deemed to be a method of improvement of search engine response by means of understanding of user intent as well as of the search terms in the context of the document space. If we take into account advances in natural language research, we easily guess that there is virtually no chance to realize the goal of semantic search, formulated in this way, in near future. Computers have no chance to understand semantics of textual messages as they have no "experience" with the reality surrounding us humans. Access to semantics of real world appears to be a remote goal.

Therefore in our research project NEKST we reformulated in a significant way the task of semantic analysis of Internet documents by understanding the task in an operational way. Instead of trying to pretend that the machine understands the meaning of the text, we use the fact that both the information sender and the recipient are human beings. Hence not the search engine but the man has to understand the text, and the search engine only supports him in this understanding. This support has the form of so-called semantic transformations on the text which on the one hand enrich the text with new features extending search characteristics and on the other hand may move the text to other space than the document space that is into the space of objects the documents are about.

So the semantic transformation means such a transformation of the document and/or query content that allows for traditional document search via a semantically related query [4,14].

Within the system NEKST the following types of semantic transformations have been implemented:

- user suggestions [22],
- substitution with synonyms, hypernyms, hyponyms and other related concepts,
- concept disambiguation [3],
- document categorization [3],
- personalized PageRank [15],
- cluster analysis and assignment of cluster keywords to documents [2],
- explicit separation of document cluster and document search,
- extraction of named entities and relations between them [23],
- diversification of responses to queries,
- dynamic summarizing [13], and
- identification and classification of harmful contents.

Of course English language there exists a significant body of research on the above-mentioned topic, but with our system we demonstrated that the concepts are also implementable for non-English and in particular for highly inflectional languages with flexible grammar like Polish.

If you take the semantic transformation view then it is obvious that you need all the traditional mechanisms of a search engine also under semantic search, including the ranking mechanisms.

Note that traditional PageRank itself is a career of a (limited) amount of semantic information. One usually assumes that a link is added to a page with some semantic relation to the pointed page in mind. But various variants of PageRank are considered in conjunction with semantic search because the links are not perfect semantic relation indicators. So one variant, implemented also by us, is the TrustRank [11] where you may give more weight to one outgoing link and less to another (e.g. based on textual similarity between the pointing and the pointed page). Still another approach is to create a kind of "topic sensitive" PageRank [12]. You may split the collection into a set of rough categories and for each of these categories one can compute a separate PageRank. If one recognizes that a query belongs to a particular topic, the personalized PageRank may be used. It may, however, happen that one query is related to two or more categories. An efficient way of computing personalized PageRank for a mixture of categories is proposed in [15].

4 Actuality Measurement

As mentioned the actuality tells us whether the Web page contains up-to-date information or was accurate in the past only.

Though we cannot confront the web content with the state of the real world, we can nonetheless check whether or not the creator of the Web page cares for its content.

Actuality can be measured by tracking the activity of web page authors. One can use the measures of deadrank or refresh rate, etc.

In the old days, to verify the actuality of a web page, one could query the date of the last update, but today many servers update artificially the creation date or the pages are dynamically generated. However, local servers are not able to overcome the dynamic nature of the Internet. Due to the frequent deletion of web pages, one can discover not maintained pages by looking at their dead links (the author did not notice disappearance of pages).

This fact gave rise to the DeadRank. Deadrank is based on the PageRank on the reverse Web graph, though it is slightly different from badrank. One thinks about the DeadRank as the probability of visiting the page by a random walker starting his journey on non-existent web pages and jumping back to them upon being bored. Not maintained pages will have high Deadrank. Of course, the problem here is to obtain information about the pages that do not exist. Usually, we can learn about the non-existence from a suitable server error, but the server can also cheat, redirecting to another page, or creating one on the fly. One can compensate for this effect by investigating what answer is given by the server if you specify an address that you are sure does not exist. Comparison of responses will give us an indication of whether the page exists or it is an artifact of the server.

Refresh rates of Web pages are also subject of manipulation by servers nowadays, but there were technologies developed for search engines guiding the visit frequency of spiders depending on some measures of refresh rate. The article [6] proposed the so-called temporal freshness, consisting of a mixture of (1) content freshness (PF), expressing updates of the page as such and (2) in-link freshness (INF), reflecting updating of the links pointing to the page. At the time i, these quantities are:

$$PFt_i(p) = \sum_{j<i} e^{(j-1)*a*\Delta t} \sum_k w_k * C(j,k,p)$$

$$InFt_i(p) = \sum_{j<i} e^{(j-1)*a*\Delta t} \sum_l w_l * C(j,l,p)$$

where w_k and w_l are weights of activities of type k on the Web page and activities of type l on in-links of the Web page resp. $C(j,k,p)$ and $C(j,k,p)$ tell if the respective activity appeared or not on the web page at time point j. Based on these data the quantity TFC (temporal correlation freshness) is defined, which is the ratio of a temporal correlation between the measurements of PF and INF. Intuition for this measurement is as follows: the site is "fresh" if its changes are seen by the links. These measurements assume that we have a historical copies of pages available (for example, InternetArchive).

5 Measurement of intelligibility

The concept of intelligibility specifies whether or not the information can be decoded by the receiver / mirrors what the sender wanted to send.

Intelligibility can be evaluated, e.g. by examining whether you can build profile of customers visiting the given page from the click stream and if the prevailed clicking behavior on the page corresponds to the obtained profile.

Other methods rely on an assessment of "the literary" quality. One examines the length of the words used on the page in terms of the number of letters and syllables, sentence length counted in words, syllables and/or letters. One also examines the variation of the words (the ratio of the number of words used to words in general, the cumulative distribution function of the words used etc.). Out of these statistics some complexity measures are derived.

One of the best known methods in this area is called Flesch Reading Ease index

$$FRi = 206.876 - 1.015ASL - 84.6ASW$$

where ASL is the average number of syllables per word, and ASW - the average number of words in a sentence[2].

More sophisticated methods can utilize the available semantic resources, including the classification of the conceptual difficulties of individual words. Usually lists of difficult words are difficult to obtain, but lists of easy words are easier to get.

Although measures related directly or indirectly to the length of the sentences are fairly good indicator of clarity, in case of hypertext documents on the internet we have to handle the problems of punctuation - it is neglected and thus the easiness may be underestimated.

But it turns out that a significant part of the easy pages point generally to the easy (easy to understand) ones, while the difficult pages point to difficult ones. Therefore, for example, one can use so-called TrustRank method [11] (an analogue of PageRank in social networks) to assess the page easiness based on prior assessment by humans of a set of seed pages.

6 Measuring availability

Availability means whether or not the information was sent by the sender.

The availability of information on the Internet can be tested by determining whether the page is a search result for search for meaningful question asked. But for the modern Internet technology the detection of availability can be much more sophisticated. First, we can ask what is the probability that a random wanderer will go to the respective information (website / web page). This concept is close to the so-called PageRank. Second, we consider the wanderer applying

[2] See `http://www.editcentral.com/gwt1/EditCentral.html` for alternative measures like Automated readability index, Flesch-Kincaid grade level, Coleman-Liau index, Gunning fog, SMOG index.

the backspace key, which leads to a measure of RBS. Thirdly, the wanderer can look for certain keywords, characterizing the respective information then we will have to deal with so-called Query-Dependent PageRank. Finally, instead of asking about the probability we can think about the ranking for the mentioned measurement in general or for a given query, consisting of one, two or three words (the minimum distance from the top ranking after these words / phrases).

We also have to take into account the technical possibilities of the presentation and the ability of human perception. From a technical point of view, it is essential to tell the format of information (standard HTML, PDF, or one of over 300 different formats of text documents, which are used to store information) and match it against the palette of formats supported by user's web browser. The issue is further extended by the problems of active sites, using JavaScript, Basic or applets and information presented in the form of files. Here we can calculate the probability of format compatibility between the source and the mouth of information.

The so-called Web Accessibility Guidelines[3], developed by W3C, indicate further need for considering the latest standards of presentation (e.g. CSS in conjunction with HTML for ornamental purposes), the usage of explanatory text, the usage of non-textual elements (graphics, sound, Flash), as well as alternatives to or features of HTML, which may be not supported by one's browser (frames, iframe, embed, etc.). Based on the use / non-utilization of such features, software has been developed such as Valet[4], which counts the number of violations of these formalist readings in the form of a linear model that presents a measure of the availability of information (see [21]).

Moreover, nowadays there is a tendency to restrict access to the Internet. Under the new legislative proposals such as the U.S. availability of information on the Internet regulation[5] is the future depend on the wealth of internet users. ISPs can freely censor information available. Therefore, the measurement will be available not just answer "yes / no", but the function of the funds available to the surfer, as well as the selection of service providers by issuing and seeking the document.

7 Concluding Remarks

Over six decades of rapid development of computer science have passed since the seminal paper of Shannon, who dared to ask how much information we can pass through an information channel. By asking this, he implied that information is an objective reality that needs to be quantified somehow. It has to be regretted that over the years one has forgotten that information is not a product of computer science and cannot be arbitrarily defined. With the advent of Web technologies it became again an acute problem get an in-depth understanding of nature and properties of information (as an immaterial dimension of the real world)

[3] http://www.w3.org/WAI/WCAG20/quickref/

[4] http://valet.webthing.com/access/url.html

[5] http://www.fcc.gov/openinternet

and how they should be measured in order to provide search engine users with most valuable output. To have a valuable output we must establish the value of each piece of information. This is somehow contrary to the current philosophical trends that hamper the actual reflection on the nature of information.

We have tried to demonstrate in this presentation why it is necessary to go beyond Shannon model, why it is important to go beyond statistical level to appropriately capture the value of the information. We have pointed to the five dimensions of information evaluation: quality, utility, actuality, intelligibility and availability. Last not least we have demonstrated that with current tools or ones available in near future it is possible to measure the information value according to these dimensions.

Surely, we do not have a proper understanding of all aspects of information and the technical realization of some concepts is far from satisfactory. But we can nonetheless hope that in near future the search engines will be able to tell us far more about the web pages we are looking for just accelerating our search for helpful information. The search engine designers have to keep in mind that the people do not live to search but rather they search to live.

References

1. Wen bing Horng, Cheng ping Lee, and Chun wen Chen. Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering 4(3)*, pages 183–191, 2001.
2. Szymon Chojnacki and Mieczysław A. Kłopotek. Grupowanie stron w polskim internecie. In *Proceedings of Artificial Intelligence Studies Vol.2012*, pages 1–8.
3. Krzysztof Ciesielski, Piotr Borkowski, Mieczysław Kłopotek, Krzysztof Trojanowski, and Kamil Wysoki. Wikipedia-based document categorization. Lecture Notes in Computer Science vol. 7053, Springer, 2012.
4. Krzysztof Ciesielski, Dariusz Czerski, Michał Dramiński, Mieczysław A. Kłopotek, and Sławomir T.Wierzchoń. Semantic information within the BEATCA framework. *Control and Cybernetics*, 39 (2):377–400, 2010.
5. Li Cunhe and Li Ke-qiang. Hyperlink classifcation: A new approach to improve pagerank. In *DEXA '07: Proceedings of the 18th International Conference on Database and Expert Systems Applications*, pages 274–277. IEEE Computer Society,Washington, DC, USA, 2007., 2007.
6. Na Dai and Brian Davison. Capturing page freshness for web search sigir'10, july 19–23, 2010, geneva, switzerland., 2010.
7. Łukasz Dębowski. Excess entropy in natural language: Present state and perspectives. *Chaos 21, 037105*, 2011.
8. Pierpaolo Dondio and Stephen Barrett. Computational trust in web content quality: A comparative evalutation on the wikipedia project, 2007.
9. B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, DUX '03, pages 1–15, New York, NY, USA, 2003. ACM.
10. Werner Gitt. *In the Beginning was Information (Am Anfang war die Information)*. New Leaf Publishing Group, (2006 edition), 1997.

11. Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 576–587. VLDB Endowment, 2004.
12. Taher H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15(4):784–796, 2003.
13. Mieczysław Kłopotek, Sławomir Wierzchoń, Krzysztof Ciesielski, Michał Dramiński, and Dariusz Czerski. *Conceptual Maps of Document Collections in Internet and Intranet. Coping with the Technological Challenge.* IPI PAN Publishing House, Warszawa,Poland, 2007. 139 pages.
14. Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, Krzysztof Ciesielski, Dariusz Czerski, and Michał Dramiński. Towards the notion of typical documents in large collections of documents. In Qingyu Zhang, Richard S. Segall, and Mei Cao, editors, *Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications.*, pages 1–18. IGI-Global, 2011.
15. Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, Dariusz Czerski, Krzysztof Ciesielski, and Michał Dramiński. A calculus for personalized PageRank. In *Proc. of IIS 2013, LNCS 7912*, pages 212–219. Springer Verlag, Heidelberg, 2013.
16. Tamara G. Kolda and Michael J. Procopio. Generalized badrank with graduated trust. Technical Report SAND2009-6670, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, October 2009.
17. Xin Li and Shawmin Lei. Block-based segmentation and adaptive coding for visually lossless compression of scanned documents. In *Proc. ICIP, VOL. III, PP*, pages 450–453, 2001.
18. Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 83–92, New York, NY, USA, 2006. ACM.
19. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
20. Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.
21. Francisco Montero Simarro, Pascual González, María Dolores Lozano, and J. Vanderdonckt. Quality models for automated evaluation of web sites usability and accessibility. In *International COS T294 Workshop on User Interface Quality Model*, 2005.
22. Marcin Sydow, Krzysztof Ciesielski, and Jakub Wajda. Introducing diversity to log-based query suggestions to deal with underspecified user queries. In Pascal Bouvry, Mieczyslaw Klopotek, Franck Leprévost, Malgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybinski, editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 251–264. Springer Berlin / Heidelberg, 2012. 10.1007/978-3-642-25261-7-20.
23. Alina Wróblewska and Marcin Sydow. Dependency-based extraction of entity-relationship triples from polish open-domain texts. volume 7(30), pages 61–70. Publ. House of University of Natural Sciences and Humanities, Siedlce, 2012.
24. Baoning Wu and Brian D. Davison. Identifying link farm spam pages. In *Proceedings of the 14th International World Wide Web Conference*, pages 820–829. ACM Press, 2005.