
Normalized Medline Distance and Its Utilization in Context-aware Life Science Literature Search

Yan Wang¹, Cong Wang¹, Yi Zeng¹, Zhisheng Huang², Vassil Momtchev³, Bo Andersson⁴, Xu Ren¹, Ning Zhong^{1,5}

¹ International WIC Institute, Beijing University of Technology, Beijing, China

² Knowledge Representation and Reasoning Group, Vrije University Amsterdam, the Netherlands

³ Ontotext AD, Sirma Group, Bulgaria

⁴ AstraZeneca R&D, Sweden

⁵ Department of Life Science and Informatics, Maebashi Institute of Technology, Japan

wang.yan@emails.bjut.edu.cn, joshua.cong@gmail.com, yizeng@bjut.edu.cn, huang@cs.vu.nl, vassil.momtchev@ontotext.com, Bo.H.Andersson@astrazeneca.com, zhong@maebashi-it.ac.jp

Abstract: When facing great volume of query results while users are searching literatures on the Web, we propose to refine the search process by using user interests. We analyze user interests and calculate semantic similarity among those interest terms to fulfill query refinement. Traditional general purpose similarity measures may not always fit a domain specific context. In this paper, under the context of medical literature search on the Web, based on the biomedical literature knowledge source “Medline”, we propose a novel similarity method—Normalized Medline Distance, which could deal with some problems of existing similarity measures and reflect relevance between medical terms more reasonably. Based on the proposed measure method, more accurate user interest description could be acquired. Based on the ranked user interests, when users try to search literatures by keywords, the most relevant results which are related to their current interests and the query input would be ranked to the top.

Key words: semantic similarity, query refinement, user interest, context-aware search

Introduction

With the fast growing life science literatures on the Web, it is difficult for scientists who want to look up some information categorized in a large ontology and formulate a query related to their interests or some of its relationships. One very typical scenario is that when people query on a large-scale literature database, they always find themselves buried in countless results.

One of the most widely applied solutions to deal with such problems is to rank the results of a query based on semantic similarity between terms in order that the system would return a set of semantically relevant results [1]. However, most similarity measures could not be well used in domain specific literature searching system. Some of them do not include enough domain dependent terms; others are not precise or high time-consuming. For these reasons, when facing domain

specific large scale knowledge processing, we emphasize that domain dependent knowledge bases are needed to support these kinds of professional search tasks. In our study, we choose a life science literature dataset—Medline (an bibliographic medical literature dataset provided by the United States National Library of Medicine) as the knowledge source to query on. According to the knowledge base, we bring up a similarity method — Normalized Medline Distance (NMD) — to handle with such domain specific situations.

In [2], we propose to refine large-scale search by user interests. According to our previous proposed user retained interest model [2], we extract researchers’ retained interests and use those interests to refine vague, keyword-based queries on Linked Life Data (LLD)¹.

As a step forward, we re-rank the extracted user

¹ <http://www.linkedlifedata.com/>

interests based on the proposed NMD measure which help to produce more precise interest description considering semantic similarities. Finally, we develop the Context-aware Linked Life Data Search Engine based on the re-ranked user interests. The experimental results show that the system could return the most relevant results which are related to the query inputs and users' recent interests.

This paper is organized as follows: Section 1 introduces the user retained interest model that we propose to extract users' explicit interests. Section 2 introduces the Normalized Medline Distance (NMD) that we propose for domain specific semantic similarity measures based on the Medline dataset. Interests similarity and interests re-ranking are investigated based on the NMD measure. Section 3 utilizes the proposed methods for literature search refinement. Section 4 concludes the paper by highlighting major contributions and briefly introducing future works.

1 User Retained Interest Model

For most researchers, their interests are shifting all the time, therefore, not only the numbers of the published articles could reflect their interests, but also the published time of those articles make an important role to calculate the interests' weights. For example, if it has been a relatively long time that someone publishes on a topic, the possibility for him/her to come back to it is very small. However, researchers who publish articles recently and continue to do research in this topic, it is rational to believe that he/she is enormously interested in it. Therefore, inspired by the forgetting mechanism, we developed users' retained interest model which could describe such phenomenon and quantitatively evaluate users' retained interests.

1.1 Definition

User interests can be divided into implicit interests and explicit interests. Implicit interests are kept in users mind and they are hard to be acquired. Explicit interests appear somewhere explicitly (such as previous publications can show some of the authors' research interests), and they can be tracked and quantitatively studied. In this paper, we focus on quantitative analysis of explicit interests. We consider

users' previous publication as the source to extract users' interests.

Firstly, we define Cumulative Interest Value (as in function 1.1) which would describe how many articles on specific topic that a user has published. All the interest keywords we choose are MESH (Medical Subject Headings) terms from Medline dataset (N-triple format). MESH terms are designed for indexing articles. They also correspond with the concepts in Medline ontology—Mesh Tree. Function 1.1-1.3 are introduced in [2,8].

We consider an interest item of a researcher as a keyword (i.e., a MESH term) which is used to describe a publication. Thus, the interests of a researcher are composed of a set of MESH terms. We use $t(i)$ to denote the i -th interest item of a researcher. The Cumulative Interest Value, denoted as $CI(t(i),n)$, is used to denote the total appeared times of the interest $t(i)$ in the n time intervals:

$$CI(t(i), n) = \sum_{j=1}^n y_{t(i),j}, \quad (1.1)$$

where $y_{t(i),j}$ denotes the total number of articles that a researcher has published with respect to the interest item $t(i)$ during the time interval j . There are n time intervals in all.

However, interests may be shifting all the time. For example, if someone published many articles around 1990s, but does not publish anything about this topic recently, it seems that the author has "forgotten" it. The above cumulative interest model cannot reflect the forgetting mechanism in the shifting process.

The process on the loss of interests is very similar to memory mechanism. According to cognitive psychology, forgetting mechanism of memory could be described as an exponential curve or a power curve^[10]. So, based on the forgetting model, we developed two retained interest model to quantitatively describe users' retained interests. Hence, we define user retained interest as follows based on exponential law and power law respectively:

$$PRI(t(i), n) = \sum_{j=1}^n y_{t(i),j} \times AT_{t(i),j}^{-b}, \quad (1.2)$$

$$ERI(t(i), n) = \sum_{j=1}^n y_{t(i),j} \times Ae^{-bT_{t(i),j}}, \quad (1.3)$$

$PRI(t(i), n)$ and $ERI(t(i), n)$ are the retained interests

models that use the power function or exponential function respectively. A and b are two parameters of these two models. We need to fit them through comparing the retained interest values and cumulative interest values in the next neighbor time interval which reflect their current interests. A is used to control the difference of the calculated retained interest value and the actual current interest value, while b is used to control the forgetting speed of the specified interest. (Here, through fitting parameters by Spearman correlation and t-test, we change the value of A and b back and forth to let Spearman tend to 1 and t-test tend to 0. So, we get that A = 0.85, and b = 1.9. The Spearman correlation is 0.628 and 2-tail t-test is 0.08, which improves a lot compared to [2,8]). T denotes the time interval between the time that the article published and now (or the specified end time).

2 Interests Similarity and Their Re-ranking based on Normalized Medline Distance

2.1 Term Similarity Measures

In the study above, we did not consider the similarity between interests, which may cause the failure of ranking the most relevant query results that users need. Semantic similarity among interest terms could adjust the query further in order that system could return a class of semantically close results.

The existing similarity measures could be divided into three different kinds, namely, WordNet-based measure, information retrieval based measure and ontology based measure. However, WordNet-based measures do not include lots of field specific words for its limitness. For example, some complex domain terms such as “Hydroxyaminoquinoline” and “Nitroquinolines” are not listed in WordNet, but medline dataset has many terms of these kinds. Information retrieval based measure need to traverse the full content of an article to calculate frequency of each term which needs a lot of time and a large scale knowledge base.

Compared to the disadvantages mentioned above, ontology based similarity measures^[3, 4] have some advantages, since those ontologies are created by experts manually and which insure that they are relatively more rational. Besides, computing similarity

in the ontology is more efficient, because we just need to traverse the ontology. But those ontology based similarity measures rely on the paths (IS-A links or other kinds of relationship links) between terms, or the depth of terms in the ontology, or the numbers of ancestors, children, and neighbours of the term. So, they could not differentiate terms which have the same properties on the above. For example, the terms “boredom”, “bereavement” and “euphoria” belong to the term “emotions”. If we calculate the similarity among these three words, the answer would be same.

Rudi and Paul developed Google Similarity Distance^[5] (function 2.1), which does not require domain specific knowledge base (Google serve as the general purpose KB here), and does not have the problems which would happen in WordNet, because you almost can find everything in Google.

$$Sim_{NGD}(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log N - \min(\log f(x), \log f(y))}, \quad (2.1)$$

f(x) denotes the number of pages containing term x. f(x, y) denotes the number of pages containing both term x and y, N denotes the number of Google indexing. According to [11], in this paper, we adopt $N = 10^{13}$.

However, there are many disadvantages for this method. Firstly, we must recognize that the similarity calculated by Google Similarity Distance has quite large noise, since it includes Web pages all over the world and many of them are not about science and technology but just general articles. Secondly, since the number of indexed Web pages is growing very quickly and Google does not provide the total indexing number, we just predict that number roughly which results in big noise. Thirdly, the similarity value calculated by Google Similarity Distance could not be normalized to [0, 1]. It is the Google internal searching mechanism that causes the indexing number that contains both two terms might be larger than the sum of the indexing numbers that contain two single term separately. This is against the basic rule of set theory. As a result, we would not know the maximum of similarity value and consequently could not normalize it. What's more tough is that Google Similarity Distance Algorithm spends too much time, since the process itself needs to download the contents of the pages from Google servers three times and to extract the index number from them. Most of the time is spent on data

transmission through httpget.

2.2 Normalized Medline Distance

Since the nature of Google Distance is from Kolmogorov complexity and Normalized Information Distance^[5,6,7], which are both abstract methods, in the specific field of Medical Science data, we propose Normalized Medline Distance (as shown in function 2.2) based on the Medline dataset. This measure could handle most of the above problems.

$$Sim_{NMD}(x, y) = \frac{\max(\log p(x), \log p(y)) - \log p(x, y)}{\text{Log}M - \min(\log p(x), \log p(y))} \quad (2.2)$$

where x and y represent two different interests respectively. $p(x)$ is the number of articles in Medline whose MESH (a comprehensive controlled vocabulary for the purpose of indexing literatures of life science) contains term x , $p(y)$ is the number of articles whose MESH contains term y , $p(x,y)$ is the number of articles whose MESH contains both term x and term y , M is the number of articles in Medline. According to the dataset we use, $M = 17196759$.

As for some general interests which appeared in the MESH terms of more than 10 million articles, for example $x = \text{"humans"}$, the $Sim_{NMD}(x,y)$ would return values over 1.0 (y is an arbitrary interest term), therefore the theoretical maximum of the value should be $\max(\log p(x), \log p(y)) = \min(\log p(x), \log p(y))$, which described as Lim , and $\log p(x,y) = 0$, which means the maximum of NMD similarity can be acquired by function 2.3.

$$Max = \frac{Lim}{\text{Log}M - Lim} \quad (Lim = \max, \min(\log p(x), \log p(y))) \quad (2.3)$$

Here $Lim = \log 10358217$, and 10358217 is the number of papers that contain the term "Humans". So, all NMD values should be divided by this maximum, in order that the range of the NMD could be in $[0, 1]$.

Here we discuss some properties of NMD similarity. NMD values are always nonnegative and $NMD(x, x) = 0$, however the equation $NMD(x,y) = 0$ does not mean $x = y$. It only indicate that $p(x) = p(y) = p(x,y)$. NMD has symmetric property, say, for every term x and term y we have $NMD(x, y) = NMD(y, x)$. However, NMD does not satisfy triangle inequality.

In addition, we generate 5000 MESH term pairs randomly and compute their similarity values using

NMD measure. From Figure 1, we could see that those values tend to normal distribution, except some points are close to the range of $[0.8, 1.0]$. In Figure 2, the more points tend to the line, the closer NMD distribution tends to normal distribution. We could see only some points that are close to $[0.8, 1.0]$ do not tend to the standard line. This experiments are designed to do t-test for evaluation which we will introduce after.

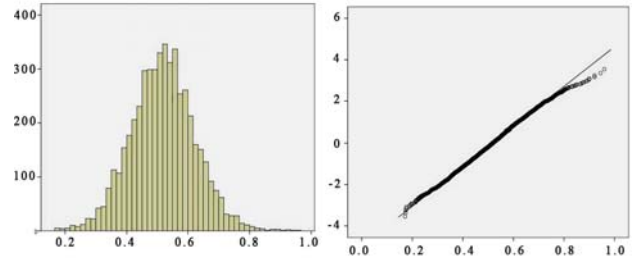


Figure 1.NMD histogram Figure 2.NMD Q-Q diagram

2.3 NMD-based Interests Reranking

In [2], we propose to refine vague/incomplete queries based on users' retained interests. Hence, the order of the interests is very important. Considering from the semantic similarity perspective, NMD values among these interests would further revise the order. For every semantically related terms x and y , $rank(x)$ and $rank(y)$ are their order values in the interests sequence respectively. After calculating NMD values, the orders will be changed to $rank'(x)$ and $rank'(y)$, and the principle could be described as follows^[8]:

$$rank'(x) = \begin{cases} rank(x), & rank'(y) = rank(x) + 1 \quad (rank(x) > rank(y)) \\ rank(y) + 1, & rank'(y) = rank(y) \quad (rank(y) > rank(x)) \end{cases}$$

It means that when term x and term y are semantically very close to each other, the term which rank backward could be updated to be close to the other one. In this way, users would get the most interesting results first and other results that contain semantically relevant terms will be ranked after them.

In our system, we assume two terms are relevant if their NMD value is less than 0.21. We randomly generate 2,500 word pairs and select 90 of them to let the experts from AstraZeneca (One of the world's leading biopharmaceutical companies located in Sweden) do similarity evaluation. The selection of the 90 pairs is based on the co-occurrence in Google to

eliminate some pairs that are not applicable. (The pairs are chosen if the index number of two words is over 50,000, otherwise the experts might have no idea about the strange pairs.). The experts are required to give a score to each pair from the set of values {1.0, 0.8, 0.6, 0.4, 0.2, 0}, among which 1.0 is the lowest relevant and 0 is the highest relevant. We choose the pairs if their experts' evaluation is lower than 0.4, which means they are relatively relevant in experts' points of view. Then, we compute the average NMD value of these relevant pairs, which is 0.21. We choose this value as our threshold for interests reranking.

2.4 Query Results Number Decision by NMD

In the MEDLINE dataset, there are several semantic abstract MESH terms, such as "Humans", "Adult" or "Female". According to our statistics, there are 17196759 articles over all in the MEDLINE dataset that we use, and there are 10358217 articles which contains the MESH term "Humans". So, after extracting user interests, more than half user interest lists include "Humans", which might lead to Triviality.

For simplicity, we could consider the terms that have top frequency as abstract interests. However, some experts might especially focus on female area. So, this method will result in unfairness for some specific experts. While, in a top 9 interest list, the more terms that a term is semantically close with, the more general, or say abstract, that term is. Based on this strategy, we set the number of results of the most abstract terms to be the lowest. On the contrary, the system will return highest number of results for the most concrete terms. Hence, this strategy could be described as follows:

$$res_num(x) = N \times (S - sim_num(x, X)), \quad (2.4)$$

$res_num(x)$ denotes the number of returned results, N denotes a constant for adjusting the largest return number. S denotes the number of interest terms in the user interest ranking list (In this study, we consider $S = 9$). $sim_num(x, X)$ denotes the number of terms that is semantically close to x in set X (here X is the user interest list).

For example, in an interest set, which includes the term "humans", the $sim_num(x, X)$ of "humans" is always 8, which indicates that the co-occurrence frequency of "humans" with other words is too high.

Then, the returned number would be $N \times (9 - 8)$. So, "humans" would be still ranked at the top, but we do not provide too many query results that contains "humans". The system would return more results for the interests which have not much relevance with other words.

3 Experiments and Utilization in Search Refinement

3.1 NMD Evaluation

NMD evaluation is designed to figure out whether our measure could reflect the similarity between words as domain experts think. We evaluate different similarity measures by correlating their similarity values with the values evaluated by judgments of experts on a given MESH terms pairs. We adopt the experts values provided in [4]. Correlation was computed using the Pearson correlation function (see function 3.1).

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sigma_x\sigma_y} \quad (3.1)$$

X is the values array of 30 terms pairs calculated by different similarity measures, Y is the values array of human judgements, \bar{X} , \bar{Y} and $\sigma_x\sigma_y$ are their means and standard deviations respectively.

We implemented different measures that are mentioned in [4,5]. SP, WP, LC, Mao and Li measures are based on paths between words, not using the frequency of a word in a dataset. Lin, Res, J measures are based on the frequency of each words, and their subsumer's frequency. J, BK, K measures are based on the numbers of ancestor nodes. So, all of the above measures do not use the co-occurrence frequency of two words. Table 3.1 provides a comparative study on the correlation of these measures.

Table 3.1 Correlation of similarity measures

Measure	SP	Li	Mao	LnC
Correlation	0.6757	0.6742	0.4070	0.6420
Measure	WnP	Lin	Res	BF
Correlation	0.6729	0.7400	0.7102	0.6884
Measure	Jiang	Knappe	NGD	NMD
Correlation	0.7299	0.7540	0.7886	0.792

It is easy to see that, our NMD measure get the highest correlation amongs these similarity measures. We believe two arrays are significantly relevant in

general, if Pearson correlation is over 0.8. However, NMD measure does not meet that condition. The reason might be that the experts' evaluated values are based on their sense about which pairs are more semantic closed, but not on which pairs are more medically relevant. This is to say, term A and term B perhaps have totally different meanings, but might be two symptoms or two treatment of one disease, or they are complications.

Since NMD function tends to normal distribution as we mentioned before, we assume the scores distribution obtained from experts also follow normal distribution. Therefore, we analyze the difference between NMD scores array and experts' scores array by t-test. The significance value 0.995 shows that NMD scores is statistically very relevant with experts' scores.

Besides, as we mentioned above, we randomly generate 2,500 word pairs and select 90 of them to let experts do the similarity evaluation. We compute the Pearson correlation value for these NMD and the experts' evaluation values, which is 0.736. While the correlation value for NGD and the experts' evaluation values is 0.531. It also indicates that NMD measure gets higher correlation on the random pairs experiment. In addition, the average values of NGD, NMD and Expert Evaluation arrays are 0.289, 0.390 and 0.590 respectively, which shows that compared to NGD, the value range of NMD is closer with Expert Evaluation. Hence in this specific domain, our measure could reflect the similarity between terms very well.

3.2 Retained Interests Extraction and Evaluation

We select 224 researchers who published around 300 articles in the MEDLINE dataset as samples in order to decide whether our method would reflect users' real interests. The reason why we choose them is that these authors' names are relatively unique, which means there would not be more than one people who have the same name. In addition, the number of articles that these researchers published is relatively high. Consequently we could use them as typical samples.

We obtain top 9 PRI values of these authors from 2001 to 2008, which means every author has 8 arrays and each of them contains 9 interests. We use PRI to predict the number of publications that is related to a specific interest in the next time interval (in this paper,

the time interval is year).

We separate these authors into 2 groups randomly. Each of them contains 112 authors respectively in order to show that the selection would not influence the effect of retained interest model.

Table 3.1 presents the fitting parameters of different groups. A and b are two constants, rho is the spearman correlation value. Table 3.2 presents the accuracy of retained interest model, 112-1 and 112-2 are the two groups that are separated from 224 authors.

Table 3.1 Fitting parameters of different groups

	112-1	112-2	224
A	0.769	0.714	0.85
b	1.671	1.533	1.9
rho	0.628	0.624	0.68
t-test	0.07	0.08	0.08

Table 3.2 Accuracy of retained interest model

	112-1	112-2	224
<i>Correctly predicted interests / the total number of interests</i>	<i>Accuracy (%)</i>		
0/9	0.095	0.095	0.095
1/9	0.019375	0.019375	0.019375
2/9	0.01375	0.01375	0.01375
3/9	0.028125	0.028125	0.028125
4/9	0.026875	0.02625	0.026875
5/9	0.06125	0.0625	0.0625
6/9	0.13	0.12875	0.135
7/9	0.255625	0.25625	0.255
8/9	0.27875	0.27875	0.271875
9/9	0.09125	0.09125	0.0925

From Table 3.1 and Table 3.2, we could see that parameter A and b would change slightly in different groups, but their corresponding rho and t-test values tend to be the same. Moreover, the accuracy of different group is almost identical as well. So, there is no difference among different groups using retained interest model, which means the model would not be affected by the types of users we choose.

From Table 3.2, we could see that the predicted accuracy is high. Over 82% of the prediction can match at least 5 interests among the top 9 interests. So, to some extent, the prediction based on interest model

is very relevant with the actual publication numbers. In addition, we did the same experiment on DBLP dataset. We get over 50% of the prediction matching. Hence, retained interests model could be widely used in any datasets of any fields.

3.3 Implementation of Context-aware LLD Search

Based on the user retained interest model and NMD-based re-ranking strategy, we extract Medical science researchers' interests (in the MEDLINE dataset) and reorganize their ranking, we use the acquired interests to add these interests terms in the previous keyword-based queries, finally we adjust the number of results that the system returns for each interest term. The system is named as Context-aware Linked Life Data Search Engine². It is developed on top of the Linked Life Data Search Engine¹. (As shown in Figure 3)



Figure 3. A Screenshot of the Context-aware Linked Life Data Search Engine

When users attempt to search literatures in the MEDLINE dataset, they will get the most relevant results that are related to their research interests and reasonable number of articles related to their current interests. The query results can be accessed online².

4 Conclusion

The major contributions of this paper are the proposed Normalized Medline Distance (NMD) measure and its application in context-aware literature search. We need to emphasize that although the NMD measure is

designed for the life science and medical domain, the method itself is applicable to other domains given the condition that there is a similar domain knowledge source for the specific domain.

The NMD similarity measure has many advantages compared to other measures, especially its correlation with experts' evaluation is higher. It could be used in many applications, such as literature clustering. Here, we just implement the NMD measure to rerank interests and adjust the number of returns. Users could acquire more relevant results to their interests than traditional search system, which could be considered as an approach for the goal of Semantic Web search.

For interests-based query refinement, the models that we use to acquire user interests focus on explicit interests, which means that if someone does not have a previous record on their interests (such as publication list, visiting query logs), their recent interests cannot be acquired. The models that could find users' implicit interests need to be investigated in our future work.

Acknowledgement

This study is supported by the European Commission under the 7th framework programme, Large Knowledge Collider (LarKC) Project (FP7-215535).

References

- [1] Zaragoza H., Najork M. Web Search Relevance Ranking. Encyclopedia of Database Systems, Springer, 2009.
- [2] Zeng Y., Yao, Y.Y., Zhong, N.: DBLP-SSE : A DBLP search support engine. In: Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence. Volume 1., IEEE Computer Society (September 2009) 626-630
- [3] Zhang X., Jing L., Hu X., Ng M., and Zhou X.: A Comparative Study of Ontology Based Term Similarity Measures on Document Clustering, Lecture Notes in Computer Science 4443, 115-126, 2007.
- [4] Hliaoutakis A.: Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline. Master thesis, Technical University of Crete, Greek, 2005.
- [5] Cilibrasi R., Vitanyi, P.: The google similarity distance. IEEE Transactions on knowledge and data engineering 19(3), 370-383, 2007
- [6] Li M., Vitanyi, P.M.B.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd Ed., Springer-Verlag, New York, 1997.

² <http://www.wici-lab.org/wici/context-aware-LLD>

- [7] Bennett, C.H., Gács, P., Li, M. Vitanyi P.M.B. and Zurek W.: Information Distance, IEEE Trans. Information Theory, 44(4), 1407~1423, 1998.
- [8] Zeng Y. Unifying Knowledge Retrieval and Reasoning on Large Scale Scientific Literatures. Ph.D thesis, Beijing University of Technology, China, 2010.
- [9] Wang Y, Zeng Y, Wang C, Ren X, Qin Y, Vassil Momtchev, Zhong N and Bo Andersson. Cognitive Memory Inspired Search Refinement on Life Science Literatures. Proceedings of the 7th International Conference on Cognitive Science, 2010.
- [10] Anderson J., Schooler L.: Reflections of the environment in memory. Psychological Science 2(6), 396-408, 1991
- [11] Alopert J., Hajaj, N.: We Knew the Web was Big. The Official Google Blog. <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>, 2008.