Modeling User Knowledge from Queries: Introducing a Metric for Knowledge

Frans van der Sluis and Egon L. van den Broek

Department of Human Media Interaction
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
`f.vandersluis@utwente.nl`, `vandenbroek@acm.org`

Abstract. The user's knowledge plays a pivotal role in the usability and experience of any information system. Based on a semantic network and query logs, this paper introduces a metric for users' knowledge on a topic. The finding that people often return to several sets of closely related, well-known, topics, leading to certain concentrated, highly activated areas in the semantic network, forms the core of this metric. Tests were performed determining the knowledgeable-ness of 32,866 users on in total 8 topics, using a data set of more than 6 million queries. The tests indicate the feasibility and robustness of such a user-centered indicator.

1 Introduction

Lev Vygotsky (1896 - 1934) and Jean Piaget (1896 - 1980) were the first to indicate the importance of a user's knowledge base for interpreting information. New knowledge should not be redundant but should have sufficient reference to what the user already knows. Moreover, sufficient overlap fosters the understandability of information (Van der Sluis and Van den Broek, in press), in particular deep understanding (Kintsch, 1994); is an important determinant of interest (Schraw and Lehman, 2001); and, through both understandability and interest, the knowledge base has an indirect effect on motivation (Reeve, 1989). Concluding, it is of pivotal importance to have a model of the user's knowledge allowing to measure the distance between a topic and the user's Knowledge Model (KM).

Several types of models have been used for user (knowledge) modeling; e.g., weighted keywords, semantic networks, weighted concepts, and association rules. A core requirement is the connectedness of concepts, which makes a semantic network particularly useful for knowledge modeling. Query logs offer an unobtrusive, large source, storing basic information about the search history (e.g., queries and some click through data) for a limited amount of time (e.g., a cookie will often expire within 6 months) (Gauch et al., 2007).

The use of a query log to base a KM on assumes that what one has searched for, one has knowledge about. We will give three arguments to support this assumption: 1) The user has to know a word before being able to posit it to a search system. This is well known as the vocabulary problem, where users

have difficulty finding the right word to represent their exact information need; i.e., they are confined to the words they know (Furnas et al., 1987). 2) A query log represents the history of searches. As a user performs a query to learn new information, a query log primarily indicates that what has recently been learned. 3) Users often return to very specific domains when performing a new search (Wedig and Madani, 2006). These domains are indicative of what is familiar to the user; i.e., domains the user has at least above-average knowledge about.

Reliability is a salient problem for user profiles based on implicit sources. Logs can contain noise; e.g., due to sharing a browser with a different user. Moreover, polysemy is a structural problem for information retrieval. Polysemy refers to the need of context in order to know the intentioned meaning of a word. Any knowledge metric based on implicit sources has to deal with this uncertainty.

The remainder of this paper looks at the feasibility of measuring the amount of knowledge a user has about a certain topic. A semantic KM is introduced in Section 2, using a query log as data source. The idea that users posit many queries on closely related topics that are very well-known to them, forms the basis for the metric of knowledge presented in Section 3. Some examples of the use of the metric are shown in Section 4. Finally, the results are discussed, reflecting on the feasibility of a web system adaptive to the knowledge base of users.

2 Knowledge Model

The KM was build on WordNet version 3.0. WordNet is a collection of 117,659 related synonym sets (synsets), each consisting of words of the same meaning (Miller, 1995). As input data, an AOL query log was used (cf. Pass et al., 2006). This log consists of 21,011,340 raw queries, gathered from 657,426 unique user ID’s. As in general a large proportion of all queries is caused by a small percentage of identifiable users (Wedig and Madani, 2006), only a subset of the most active 5% of users was used. One outlier was removed, having 106,385 queries compared to 7,420 for the next highest user. The final subset consisted of 32,866 users and was responsible for 29.23% of all unique queries.

The procedure used to analyze the query logs is an iterative one, analyzing each unique query present in the (subset of the) log. For each query (q), the total set of synsets touched upon (S_q) was computed as follows:

$$\begin{aligned} S_q(q) &= \bigcup_{w \in q} S(w), \\ S(w) &= \{s \in W | w \in s\}. \end{aligned} \tag{1}$$

Here, $S(w)$ gives the set of all possible synsets s of lexical dictionary W (i.e., WordNet) related to the word w of query q . All possible lemmata for each word w were used to retrieve the synsets s .

With the analysis of the query log noise reduction was disregarded; no effort was spent to reduce the effects of polysemy. This resulted in substantial noise within the data set. On average, for the subset of the data used, a query loaded upon 13.92 synsets. The results of the analyses were stored as a large collection

Fig. 1. Average related synsets.

Fig. 2. Activation limit function a .

Fig. 3. Distance decay function d .

of (user, query, synset, part-of-speech) tuples. This collection allowed for the retrieval of the number of times a user posited a query on a specific synset $Q(s)$.

3 Similarity Metric

The distance between a certain topic and the KM indicates the amount of knowledge a user has on that topic. We exploit this by measuring the semantic distance $k(t)$ (See Equation 6) between a synset t , representing the topic, and the KM. This is done by looking at the number of activated synsets A near the topic synset t , which is inspired by the finding that people stick to certain domains for their searches (Wedig and Madani, 2006). This leads to a large number of (unique) queries in a specific domain and, accordingly, to very few queries in non-familiar domains. By looking at areas of activation, the method attends one of the core problems of user profiling: polysemy. It is expected that polysemy will lead to the random, isolated, activation of synsets and not to the activation of an area of synsets.

The number of activated synsets $A_n(t)$ is defined as:

$$A_n(t) = \#\{s \in S_n(t) | Q(s) > 0\}, \quad (2)$$

where $Q(s)$ is the previously defined number of queries per synset s (See Section 2) and S_n is the set of synsets in exactly n steps related to topic t (Gervasi and Ambriola, 2003):

$$\begin{aligned} S_0(w) &= \{s \in W | w \in s\}, \\ S_n(w) &= S_{n-1} \cup \{s \in W | r(s, s') \wedge s' \in S_{n-1}(w)\}. \end{aligned} \quad (3)$$

Here, w is the word of interest (i.e., the topic t), s represents all the synsets in lexical dictionary W (e.g., WordNet) of which the word w is part of, and $r(s, s')$ is a boolean function that indicates whether or not there is any relationship between synset s and s' . Please note that S_0 is the same as S from equation 1 and that $S_n(w)$ has a memory: it includes the synsets up to n . Following Gervasi and Ambriola (2003), hyponymy relationships are excluded from function $r(s, s')$.

The absolute count A is not directly useful for two reasons. First, the growth in synsets for higher ns is approximately exponential (cf. Figure 1). This can

Table 1. Exemplar users.

User	A_0	A_1	A_2	A_3	A_4	A_5	k
1	0	1	1	1	1	1	0.48
2	1	1	1	1	1	1	0.98
3	0	3	3	3	3	3	0.85
4	1	3	4	4	4	5	1.38

lead to very high values of $A_n(t)$ for higher ns . Second, a normalized value is preferable, allowing to estimate the limits of the function $k_N(t)$. Consequently, there is a need for a strict limiting function for the contribution of $A_n(t)$ to $k_N(t)$. Let c_1 be a constant and A_n be defined in Equation 2; then, the weighted activation a at n steps from topic t is given by:

$$a_n(t) = 1 - (c_1)^{A_n(t)}. \quad (4)$$

The constant c_1 is indicative for how many activated synsets are needed. Since lower values of $A_n(t)$ are very informative (as opposed to higher values), we choose for a value of $\frac{1}{2}$. Figure 2 illustrates this idea.

It is unlikely that a user has as much knowledge on synsets $n = 3$ steps away as on synsets $n = 1$ step away. Therefore, a decay function is needed, giving a high penalty to activation far away. We use the following exponential decay function to achieve this:

$$d(n) = (c_2)^n, \quad (5)$$

where n is the number of steps of the connection and c_2 a constant determining the distance penalty. The output of this function is illustrated in Figure 3 for $c_2 = \frac{1}{2}$ over $n = 10$ steps. The idea clearly shows from this figure. Activation in the KM close to the input synsets are given a higher weight, to such an extent that twice as much weighted activation is needed at $n = 1$, four times as much activation at $n = 2$, etc.

Combining Equations 4 and 5, the knowledge a user has on a topic synset t can be estimated by the function $k(t)$:

$$k_N(t) = \sum_{n=0}^N d(n)a_n(t). \quad (6)$$

The result is a weighted count of how many related synsets a user has previously touched upon, able to indicate how close a topic is to the user’s knowledge. The maximum of the function is 2.00, the minimum is 0. A short review of the behavior of $k_N(t)$ helps interpreting its range. Consider some hypothetical users and their unweighted activation values of $A_n(t)$, as shown in Table 1. The values of $k_N(t)$ are shown in the last column. An interpretation of the values of $A_n(t)$ is that the first user has relatively little knowledge about the topic, the second and third have reasonable knowledge, and the fourth has a lot of knowledge. Therefore, as a rule of thumb, we use the value of 0.85 as a threshold.

Fig. 4. Distribution of users over topic relatedness.

4 Proof of Concept

As a proof of concept, the KM and metric were tested on eight topics. The performed test method defines, for all users, their knowledge about the topic. This was done through the following steps. Each topic is represented by a word; subsequently, for every word, the corresponding synsets were collected. Next, for each user, $k_N(t)$ was computed. A value of $N = 5$ was used. The distribution of users over $k_N(t)$ is illustrated by a 10-bins histogram, counting the number of users per bin of $k_N(t)$, shown in Figure 4. Furthermore, the total number of “knowledgeable users” (K) for each topic was calculated by counting the number of users with $k_N(t) > 0.85$; i.e., the previously defined threshold. The values of K for each topic are shown in Table 2.

Three of the eight topics were derived from a list of topics that Wedig and Madani (2006) identified as being sticky topics: those topics that a group of users often return to. The sticky topics are: lottery, travel, and money. Figure 4(a-c) illustrates that for all three topics there is a particular user group active in the topic and a greater set of users inactive; i.e., with $k_N(t) < 0.85$. From these distributions of users, the knowledge metric can identify knowledgeable users.

Furthermore, Figure 4d shows a very uncommon word: malonylurea (Gervasi and Ambriola, 2003). This word refers to a chemical compound often used in barbiturate drugs. Being a word previously unknown to the authors of this article, the expectation was that this word will not have many users common to the word. This expectation was correct, as $K = 0$.

The topic vacation is, besides a very common word, also a somewhat peculiar word concerning its semantic relatedness. The word corresponds to two synsets

Table 2. Number of knowledgeable users per topic.

Topic	K	Topic	K	Topic	K
Health	5,411	Money	3,082	University	7,551
Lottery	3,624	Malonylurea	0	School	14,077
Travel	3,960	Vacation	4,952	University and School	4,701

that after following its relationships for $n = 3$ steps still lead to only $S_3(t) = 8$ synsets. Compared to the average, this is a very low number of synsets; see also Figure 1. However, the metric can still identify (at least part of) the knowledgeable users: $K = 4,952$; see Figure 4e).

Finally, two somewhat related words were analyzed: university (Figure 4f) and school (Figure 4g). Both are very popular topics, respectively having $K = 7,551$ and $K = 14,077$ users. Moreover, the topics were combined by taking only those users regarded as knowledgeable users (K) on both topics and averaging their values of $k_N(t)$. This was still a large number of users: $K = 4,701$; see also Figure 4h. Hence, a concatenation of topics also seems possible.

5 Discussion

Founded on the notion that people perform most of their searches on a few salient topics, this paper introduced a metric of knowledge. Queries of users were monitored and the synsets used were identified. This revealed that areas of the semantic network on which a user had posed queries before were most activated. This provided a metric of users’ knowledge on a topic. Using the metric introduced in Section 3, Section 4 showed that it is indeed feasible to give an indication of a user’s knowledge. Moreover, we showed that this indication can be founded on i) relatively limited data, ii) only a restricted KM, and iii) a noisy data set, as no effort was put into noise reduction. So, if based on a measure of the spread of activation, the possibility of “measuring knowledge” seems to be quite robust and can, thus, be applied to make information systems adaptive.

Several improvements can be made to the source, model, and metric. To start with the source, the use of queries provides limited information. This is partly due to the inherent ambiguity of the often short, keywords-based, queries. However, users also, often, only search for items that are at a certain moment in their interest. Hence, using only this source gives a skewed KM. A more elaborate approach could for example be the use of the content of a user’s home directory.

The metric itself has room for improvement as well. For example, the constants c_1 and c_2 can be further optimized. Moreover, the implementation of semantic relationship, as introduced in Section 3, can be improved. Currently, it is a simple distance measure: the n relations a synset is away from a topic. Other measures have been proposed as well; e.g., Hirst-StOnge, Resnik, Leacock-Chodorow, and Jiang-Conrath (Budanitsky and Hirst, 2006). However, not all features will be suitable for our purpose – a measure of distinctness between two

items of knowledge (or, information) is needed, which is consistent with how the user will experience it (Van der Sluis et al., in press). A similar notion holds for the threshold of .85, used to indicate when a user is knowledgeable: user tests are needed to compare the metric to how the user perceives his knowledgeable-ness.

The last argument of perceived relatedness and knowledgeable-ness is intrinsically interweaved with the used model of knowledge. For the purpose of modeling knowledge, a semantic model such as WordNet is distinct from a true reflection of knowledge. Besides the absence of named entities in WordNet, one of the most salient problems is the small number of relations needed to converge a large part of all synsets. This causes the effect seen in all examples shown in Figure 4, where almost all users tend to obtain a low relationship with the topic of interest. This is even the case for very uncommon words such as *malonylurea*. Hence, a more elaborate model can hold much more power in distinguishing between parts of the model, alleviating the need for a strict decay function; see also Figure 3.

Every KM that is based on implicit sources will be particularly successful in identifying true positives; i.e., topics on which a user has knowledge. In contrast, the identification of false negatives forms a more substantial challenge. No one data source will cover all the knowledge a user has. Every KM will also, to a lesser extent, suffer from false positives. This can occur when people simply forget about a topic or when they share their account with a different user. However, this is less of a problem when looking at the spread of activation, as this spread indicates that a user has often returned to that topic. Moreover, when using not up-to-date logs, this problem should be less prominent.

LaBerge and Samuels (1974) noted about comprehension: “*The complexity of the comprehension operation appears to be as enormous as that of thinking in general*” (p. 320). Notwithstanding, by looking at the spread of activation extracted from a user’s query history, it is feasible to infer part of the comprehensibility: the relatedness of a (new) topic to a model of the user’s knowledge. This metric will pave the way to adaptive web technology, allowing systems to directly aim for a user’s understandability, interest, and experience.

Acknowledgements

We would like to thank Claudia Hauff, Betsy van Dijk, Anton Nijholt, and Franciska de Jong for their helpful comments on this research. This work was part of the PuppyIR project, which is supported by a grant of the 7th Framework ICT Programme (FP7-ICT-2007-3) of the European Union.

Bibliography

- Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.
- Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971.
- Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). User profiles for personalized information access. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, chapter 2, pages 54–89. Springer, Berlin.
- Gervasi, V. and Ambriola, V. (2003). Quantitative assessment of textual complexity. In Merlini Barbaresi, L., editor, *Complexity in Language and Text*, pages 197–228. Pisa, Italy: Plus Pisa University Press.
- Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49(4):294 – 303.
- LaBerge, D. and Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2):293 – 323.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.
- Pass, G., Chowdhury, A., and Torgeson, C. (2006). A picture of search. In *Proc. 1st Intl. Conf. on Scalable Information Systems*. ACM Press New York, NY, USA.
- Reeve, J. (1989). The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13(2):83–103.
- Schraw, G. and Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review*, 13:23–52(30).
- Van der Sluis, F. and Van den Broek, E. L. (in press). Applying Ockham’s razor to search results: Using complexity measures in information retrieval. In *Information Interaction in Context (IiIX) Symposium*, New York, USA. ACM.
- Van der Sluis, F., Van den Broek, E. L., and Van Dijk, E. M. A. G. (in press). Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *Third International Workshop on Web Information Retrieval Support Systems August 31, 2010 (Toronto, Canada)*.
- Wedig, S. and Madani, O. (2006). A large-scale analysis of query logs for assessing personalization opportunities. In *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 742–747, New York, NY, USA. ACM.