# Current expertise location by exploiting the dynamics of knowledge

*Josef Nozicka*

*Faculty of Informatics and Statistics*

*University of Economic, Prague*

*Czech Republic*

*nozj01@vse.cz*

**Abstract:** *Systems for expertise location are either very expensive in terms of the costs of maintenance or they tend to become obsolete or incomplete during the time. This article presents a new approach to knowledge mapping/expertise location allowing reducing the costs of knowledge mapping by maintaining the accuracy of the knowledge map. The efficiency of the knowledge map is achieved by introducing the knowledge estimation measures analysing the dynamics of knowledge of company employees and their textual results of work. Finding an expert with most up-to date knowledge is supported by focusing publishing history analysis. The efficiency of proposed measures within various timeframes of publishing history is evaluated by evaluation method introduced within the article. The evaluation took place in the environment of a middle-sized software company allowing seeing directly a practical usability of the expertise location technique. The results form various implications deployment of knowledge map within the company.*

**Keywords:** Knowledge management, knowledge search, knowledge mapping, expertise location, dynamics of knowledge, publishing history exploitation

## 1. Expertise location at glance

Expertise location (or knowledge mapping which I consider for synonyms in this article) can take various forms, starting by relaxed approaches relying only on documents produced without adding any support for finding the right expert. Finding then the author of the document does not automatically imply that he is an expert. The alternative is a social network search in both old and new senses of the expression: In the old sense asking somebody who should know the knower, eventually propagating the question over a network of relationships to somebody, who knows the answer. Or, in the new sense, the search is performed using social networking software by e.g. posting the question on a Facebook profile. The process has in both cases very similar properties. It involves the attention of many people, who are interrupted from other tasks and requires their effort in case they are only mediating the answer. In some cases finding the expert this way leads to an incomplete or inoptimal result, when the right experts are not found for the lack of social connections or geographical distance.

Expert directories described e.g. in (Smith & Farquhar, 2000). Expert directories take usually form of manually (often self-) maintained profiles describing the knowledge of the employee in form of keywords or key phrases. The maintenance of this type of directory is very hard, discipline demanding. After a shorter or longer time the profile does not reflect the current state of knowledge of the owner. Not only because of the lack of effort, but also from the lack of reflection of ourselves, we all have. In literature, this problem is often described being part of tacit knowledge.

Knowledge/expertise maps described e.g. in (Busch, et al., 2001), (MITRE, 2008), (Mockus & Herbsleb, 2002), (Vivacqua, 1999), (Yu, et al., 1999) are contrary to expert directories usually automatically generated. Various document sources are indexed including documents, e-mails etc. Ontology is often used as a categorization mechanism. As in the case of other categorization mechanisms, ontologies as well tend to be stable. They do not allow formulating searches to knowledge areas not known in time of ontology construction/assignment. The process of feature extraction in case of assigning an expert to a particular knowledge reflected in sensitive documents like e-mails has to conform to some sort of document handling rules. In some cases it involves, as described in e.g. in (Jambrich, 2005), some sort of author's cooperation on at least approval of publishing areas of his knowledge to prevent sensitive knowledge to be published within the knowledge/expertise map.

## 2. Knowledge mapping model

Contrary (in some senses) to previously mentioned approaches, a knowledge mapping model described in this article

- extracts the information about knowledge from resources already available within the company, but with respecting the sensitivity of extracted information,
- presents the results in form of list of experts arranged by their level of expertise following the principle of fostering cooperation and knowledge creation as defined in (Nonaka & Takeuchi, 1995). The main aim is therefore to allow to find the most qualified and skilled persons in a company within the questioned area of expertise. In some scarce areas that requirement could mean to find somebody, who knows even anything about the subject.
- allows overview of knowledge over often large scale distributed organisation allowing to find knowers even in geographically distant areas
- is able to distinguish the search for the overall expert and search for an expert with most up to date knowledge (let's call him current expert). This article focuses on current expert search.
- Areas of expertise are not predefined. Moreover the model should allow searching expertise in emerging areas. In consequence it means as well, that the set of questions isn't predefined – that means any set of keywords could form the search phrase.
- The maintenance of the model shouldn't be the task for the already overwhelmed experts.
- The model is maintained automatically.

Other theoretical requirements for the model are described deeper in (Nožička, 2003).Technically our model has the following properties:

- Basis for expertise recognition are documents stored within various types of company document repositories and the relationship of the document to the person (usually the author) expressed in the metadata. The model of expertise is grounded in the full text index of available text sources including the metadata. The measures of knowledge estimation are built on top of the model.
- The model is domain independent – allowing to be used within any problem domain.
- The model was developed for the needs of business analytical department of a software developing company. The department has about 20 business analysts, whose daily work is to produce and update documents describing contents and functionality of various software products. Various sorts of document repositories are used (starting by shared directory structures, through project management intranets to document versioning repositories like MS Visual SourceSafe or Subversion). The basic requirement of the model to be implemented in an analogical way in any domain is that the results of the work should be available in textual form and be regularly updated and accessible to the knowledge mapping application.
- Free text queries consisting of any terms of interest are allowed. The expertise model is generated at query time allowing retrieving the always up to date information from the updated full-text index, to answer queries not known in time of document indexing. As well it allows a very prompt reaction to changes in the state of source documents within the company.

Those are the properties of the model according to higher mentioned classifications, but what is the difference that makes the model different from others?

## 3. Mining the knowledge dynamics

Generally two basic properties make the model distinct: The first is the inference of estimated knowledge from publication activity history of the author and is analysed within this section. The second one is the focussing on particular timeframe of publication history when a specific to date knowledge is looked up and is explained in the next section.

As previously mentioned, knowledge profiles are generated at runtime. At query time all documents relating to queried term and a person form a basis for analysing the level of knowledge of queried term by a particular person. In this step various properties of relevant documents are extracted: except for the authorship/other document relationship as well a score provided by the indexing machine. One of the key analysed properties in the research is the date of the document publication or last modification. The publishing/modification date allows analysed documents being considered important knowledge indicators of a person at a particular time. The core of the research described in this paper is trying to assess the level of knowledge of particular person by analysing the sequence of such knowledge indicators.

Various measures estimating the level of knowledge based on various theories of cognitive science were explored and examined in the sense of being able to arrange people according to their level of knowledge of particular area of interest. They could be classified into groups arranged in the following subchapters:

- Measures estimating expertise by regular exposure to queried subject
- Measures estimating knowledge according to learning and forgetting
- Combined measures

## 3.1 Measure estimating expertise by regular exposure to queried subject

The main indicator of a knowledge level in this type of measure is the frequency of publication of the subject by given author. If we put published documents on queried subject on a timescale and the height of the bold black column depicts individual document score assigned to document by indexing machine and normalized to level > 1 ($s_{qad}$), then this measure estimates a higher level of expertise (depicted by the green area) in case the period between publishing on the same subject is small enough. Smaller levels of expertise are estimated by grater periods between publications. The heuristics behind infers the level of expertise from the frequency of knowledge usage. The more frequent the usage, the higher probability of good knowledge:

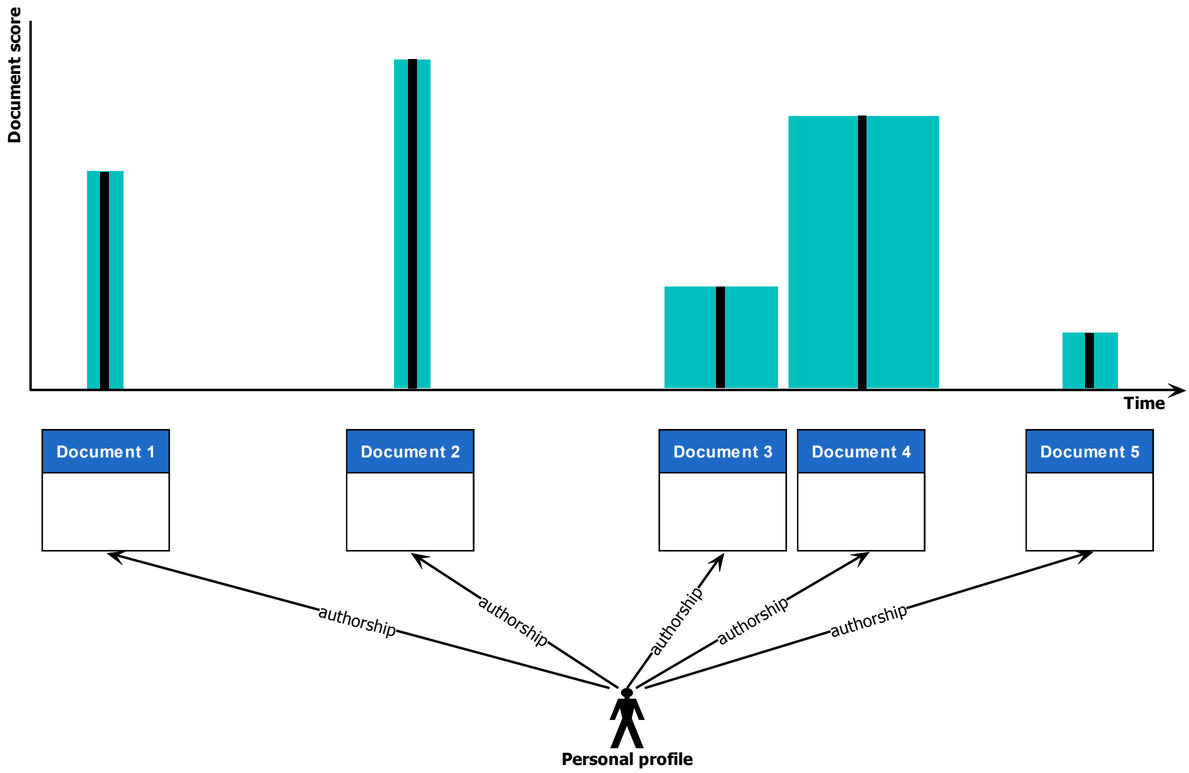


*Fig. 1: Estimating expertise by regular exposure to queried subject*

Let's have a result set (*D*) of query (*q*) of (*n*) documents indexed by (*i*) ordered by the time of document (*d*) publication (*t*) where author (*a*) mentions queried term (*q*):

$$\forall d_i \in D: t_{d_{i-1}} \leq t_{d_i} \leq t_{d_{i+1}} \qquad \textbf{(1)}$$

One of the possible instances of this heuristics sums individual document scores assigned by indexing machine ($s_{qad_i}$) for documents that are distant more than k days in time ($t_{d_{i+1}} - t_{d_i}$), whereas sums product of scores of actual document ($s_{qad_i}$) and following document ($s_{qad_{i+1}}$) for documents that are a constant (*k*) days and closer to each other in time:

$$S_{Exposure\,qa} = \sum_{i=1}^{n} \begin{cases} s_{qad_i} s_{qad_{i+1}}\ for\ t_{d_{i+1}} - t_{d_i} \leq k\ days \\ s_{qad_i}\ for\ t_{d_{i+1}} - t_{d_i} > k\ days \end{cases} \quad (2)$$

## 3.2 Measures estimating knowledge according to learning and forgetting

This type of measure is driven by the theory of learning and forgetting. Documents published at given time are considered to be indicators of knowledge of particular subject at given time. The time that follows the use or publishing of the knowledge according to knowledge theory is the time, when forgetting starts. The forgetting continues until the knowledge is demonstrated again. The curve of forgetting usually steeply decreases immediately after knowledge use and loosely decreasing thereafter. When applying this approach repeatedly, current estimated level of knowledge should be on the level marked by the intersection of green and interrupted line in the picture:

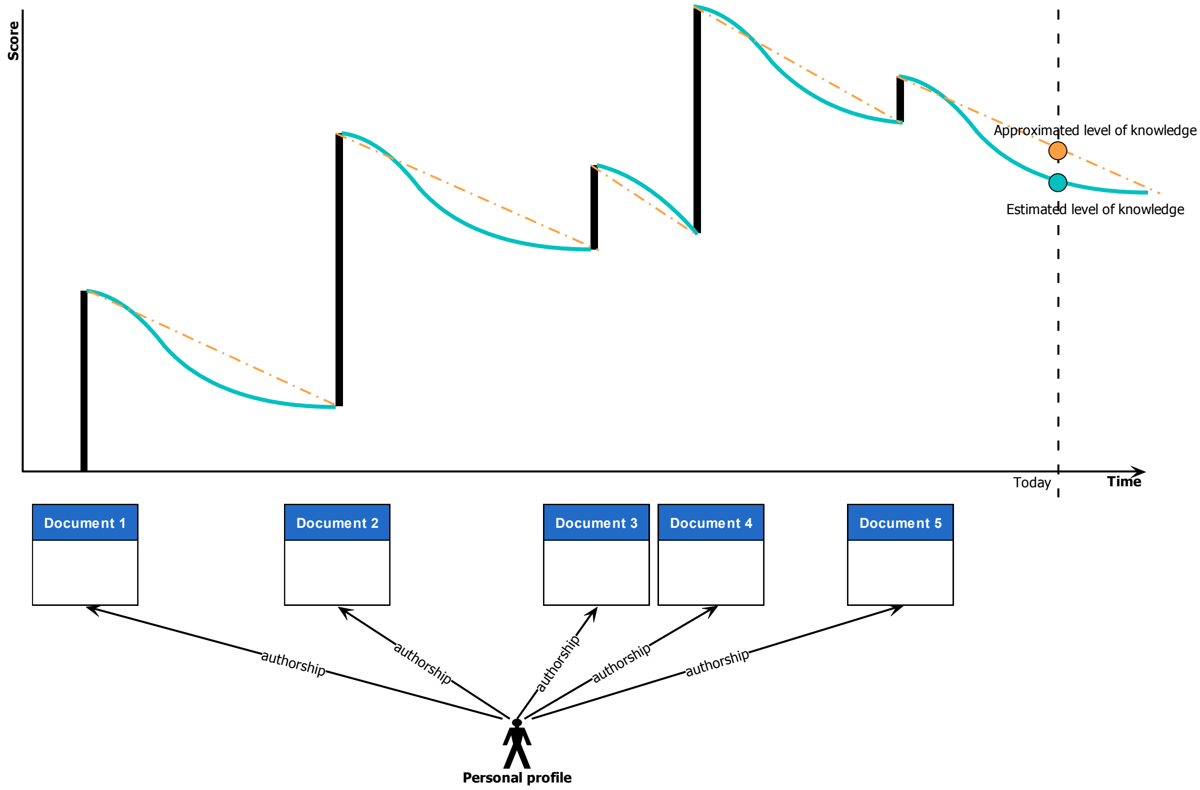


*Fig. 2: Estimating knowledge according to theory learning and forgetting*

Let's have again an ordered result set defined in **(1).** In a linear approximation indicated in the picture by dot and dashed orange line the level of knowledge (score) of an author after publishing each document ($d$) is defined as

$$S_{qa}(t_{d_i}) = \begin{cases} S_{qa}(t_{d_{i-1}}) - c(t_{d_i} - t_{d_{i-1}}) + s_{qad_i}\ for\ S_{qa}(t_{d_{i-1}}) - c(t_{d_i} - t_{d_{i-1}}) > \frac{S_{qa\,max}}{4} \\ \frac{S_{qa\,max}}{4} + s_{qad_i}\ for\ the\ rest \end{cases}, \quad (3)$$

where ($c$) is a constant describing the rate of forgetting and $S_{qa}(t_{d_1}) = s_{qad_1}$. According to learning theory introduced by Bahrick (Sternberg, 2002) the long term knowledge never drops under some minimal level, which is roughly estimated at ¼ of its maximum ($S_{qa\,max}$). $S_{qa\,max}$ is defined after every step of $S_{qa}(t_{d_i})$ counting as a maximum from current $S_{qa\,max}$ value and $S_{qa}(t_{d_i})$ for current document:

$$S_{qa\,max} = \max(S_{qa\,max}, S_{qa}(t_{d_i})) \quad (4)$$

The $S(t_{current})_{qa}$ is a function estimating the current level of knowledge of a particular author after recursively applying forgetting function to all levels of knowledge demonstrated by the author by publishing documents on particular queried subject (*q*). The value of scoring function characterizing level of knowledge in current moment (labelled $t_{current}$)is defined as:

$$S_{qa}(t_{current}) = \begin{cases} S_{qa}(t_{d_{i-1}}) - c(t_{current} - t_{d_{i-1}}) \; for \; S_{qa}(t_{d_{i-1}}) - c(t_{current} - t_{d_{i-1}}) > \frac{S_{qa\,max}}{4} \\ \frac{S_{qa\,max}}{4} \; for \; the \; rest \end{cases} \quad (5)$$

## 3.3 Combined measures

Combined measures are combining both of the previously mentioned approaches to estimate the level of knowledge. The level of knowledge demonstrated by a particular document had to be learned before. After demonstrating a particular level of knowledge, forgetting takes place. In case the publishing of documents successes in short interval, then forgetting and learning does not have a strong effect. The total score of an author is then the area under curves depicted (without linear approximations) e.g. in following picture:

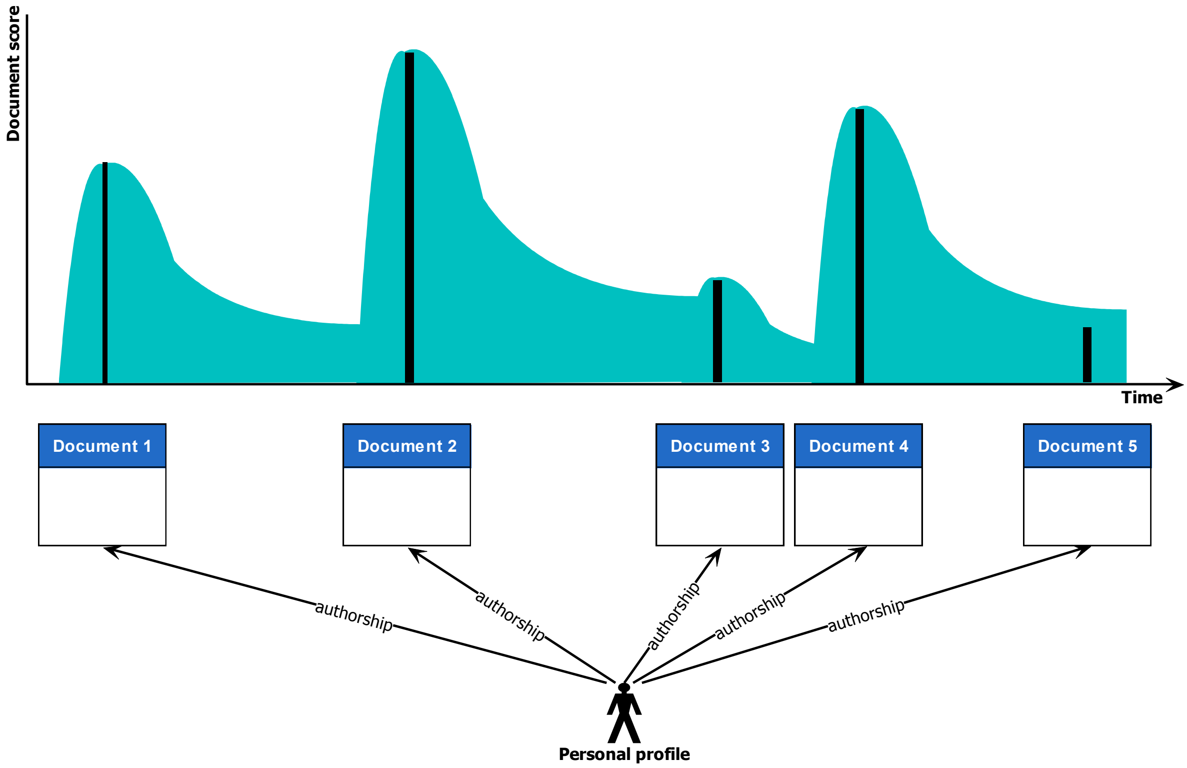


*Fig. 3: Estimating knowledge by combining forgetting and learning theory with exposure to the subject*

In terms of mathematics: Let each document is represented by its particular score $s_{qad}$ denoting the level of knowledge of given author at document publication date ($t_d$). Then the learning function ($learning_{qad}(t)$) at any time (t) is defined as an growing function having its peak in time of document publication followed by decreasing function starting in the date of document publication. Again a linear simplification has been applied in contrast to usual non-linear learning function representation. $C_l$ and $C_f$ being the linear constants for rate of learning and forgetting:

$$learning_{qad}(t) = \begin{cases} s_{qad} - C_l(t_d - t)\ for\ t_d - \frac{s_{qad}}{C_l} < t \leq t_d \\ s_{qad} - C_f(t - t_d) for\ t_d > t > \ t_d + \frac{s_{qad}}{C_f} \end{cases} \tag{6}$$

As the knowledge is naturally conceived as a non-negative variable, please note, it is defined only in non-negative boundaries.

The function $learning_{qad}(t)$ makes very minimal assumptions about knowledge of a person contributing particular document related to questioned term. The function counts with it, that for demonstrating knowledge in a document the knowledge was learned from scratch and that after publishing the document, knowledge is decreasing its level by the rate specified in $C_f$ until forgetting. That is no previous and posterior knowledge is presumed.

But this is surely not the case for experts regularly publishing documents. Therefore a function characterising knowledge by a specific author $learning_{qa}(t)$ is defined as a maximal value from any $learning_{qad}(t)$ on all documents (*d*) from a set of relevant documents (*D*) at any given time (*t*):

$$learning_{qa}(t) = \ \max\left(\forall d\ \in D\ learning_{qad}(t)\right) \tag{7}$$

The knowledge score for a particular author is then understood as a surface under the learning curve from past until current moment ($t_{current}$):

$$S_{qa} = \int_{t=0}^{t_{current}} learning_{qa}(t) \tag{8}$$

## 4. Publication history focussing

The types of the search in knowledge dependent companies could be divided roughly into two (often overlapping categories)

- Search for the best overall knowledge – the searcher is searching for somebody with the deepest knowledge of the subject. He doesn't have to be interested in latest advancements in the subject. A good example is the search for a senior employee - somebody, who is able to answer some deeper questions or somebody, who is able to give a good overview over the subject. The search for overall knowledge is discussed in (Nožička, 2012).
- Search for most up to date knowledge – or junior employee search could be used, when the the searcher is looking for somebody who has the latest knowledge in particular subject. This article focuses on this type of search.

Various possibilities, how to achieve reliable results in search for most up to date knowledge (junior employee search) were analysed:

- Knowledge estimation measures discounting older knowledge indicators in comparison to newer ones. A good example is a measure summing up the scores and dividing them by number of years between publishing date ($y_{qad_i}$) and current moment ($y_{current}$):

$$S_{DiscountSumqa} = \sum_{i=1}^{n} \frac{s_{qad_i}}{y_{current} - y_{qad_i} + 1} \tag{9}$$

- Focussing on particular newer areas of document publishing history (and omitting older ones). That means evaluating publishing history only in given timeframe. Overall, three years and one year timeframes were analysed.

## 5. Surveying knowledge for knowledge map validation

How reliable are higher mentioned measures and how do they perform in comparison? One of the key problems of the research was to identify an appropriate validation method of the model results. Various obstacles and sources of imprecision had to be taken into account: Human knowledge is hardly expressible in its whole. Tacit knowledge builds additional barrier to knowledge externalization and measurement. The only source of knowledge about knowledge was the people somehow familiar with other's knowledge and capable of a comparison between people according to their expertise. The research was conducted on an organization unit of a company small enough for the people to have a good overview of each-other's competences. The company has more than 100 people, and about half of them were during the history an expert member of the unit. Nobody of course could be considered as having full overview of the knowledge within the organization unit. Therefore as a method of objectification more people considered as having good overview were asked to take place in the research. A questionnaire survey was prepared requesting 4 respondents in 36 areas of knowledge (terms) to find (if possible) maximum 5 current experts and to align them in the order of their current expertise level. Individual results (orderings of max. 5 experts assigned by each respondent to each of the 36 areas of knowledge) were then analysed and where different orderings among respondents, were consulted again with surveyed persons to determine an ordering as close to objective as possible.

The comparison of the results of the survey to actual results of the model was performed by metrics (the term metric is used throughout the whole article to distinct them clearly from measures of knowledge estimation described in previous chapter) originating in information retrieval theory: "Precision and recall", although they had to be adapted to specific conditions of expertise search. Let's have a set of experts identified by the survey ($E_{Sq}$) and a set of experts identified by the knowledge mapping model ($E_{Mq}$). A recall function – the ability to find all the experts identified by the survey – for a queried term/area of expertise ($q$) is then defined as:

$$r(q) = \frac{|E_{Sq} \cap E_{Mq}|}{|E_{Sq}|} \quad \textbf{(10)}$$

Except for classical recall some more strict metrics (not defined here mathematically) were adopted:

- The ability to find $x$ experts identified by the survey and to rank them as the most important x experts ($r_x(q)$)
- The ability to find x experts identified by the survey and to rank them in the set of experts 2x large ($r_{2x}(q)$)
- The ability to identify the first expert in( $r_1(q)$)
- The ability to identify the first expert in first 3 found ($r_3(q)$)

Direct derivation of "precision" directly from information theory was impossible, because of the lack of reliable information about "not knowing" of particular persons. Therefore an approach based on total distance in ordering of experts in the model and in survey was adopted in measure ($d$). Another measure counted the distance only for first 5 experts identified by the model ($d_5$).

The validation was performed on knowledge mapping model within one of the divisions of middle scaled software company. The source of the modelling was the whole common analytical documentation publicly available within the company of various developed software systems. The documentation comprises of two sources – one is the directory structure shared by the software analysts to share various common documents. At the time of the evaluation it contained 12955 documents. The second one is the document repository/versioning system Microsoft Visual Source Safe (MS VSS) used to cooperative creation of analytical documents and contained 8830 documents and their versions. Both of the parts cover the complete history of documents of the analytical division (about 10 years) and are evaluated under the name „combined source" in *Fig. 4*.

## 6. The results

The following tables present the results of model evaluation. The columns show the performance of measures described earlier, the first measure being included for comparison do not take into account the history of document publication. Results of the metrics are presented in the rows in percentual form - 100% showing the absolute accuracy in comparison of the model and the survey. Evaluation of recall was performed twice. Firstly as an absolute recall – takes into account all the experts mentioned in the survey. Relative recall then operates only with authors that were named in the survey as well as contributed to the repository by at least one document – publishing authors. The recall (*r(q)*) always demonstrates the level of participation of authors within the document repository – the percentage of experts publishing within particular document repository.

The following table shows, how the model performed when it was not focused on any particular area of document history – counting the whole ten years document history of MS Visual SourceSafe repository:

*Tab. 1: Performance of the model during the whole history evaluation*

| | | | Knowledge estimation measure | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $S_{Sumqa}$ | $S_{Exposureqa}$ | $S_{DiscountSumqa}$ | $S_{qa}(t_{current})$ | $S_{qa}$ | Average | Average |
| Evaluation method | absolute recall | $r$ | 61,94% | 61,94% | 61,94% | 61,94% | 61,94% | 61,94% | 52,95% |
| | | $r_x(q)$ | 45,60% | 43,89% | 42,13% | 45,00% | 44,58% | 44,24% | |
| | | $r_{2x}(q)$ | 58,70% | 58,70% | 58,70% | 58,70% | 58,01% | 58,56% | |
| | | $r_1(q)$ | 41,67% | 44,44% | 41,67% | 38,89% | 41,67% | 41,67% | |
| | | $r_3(q)$ | 58,33% | 58,33% | 58,33% | 58,33% | 58,33% | 58,33% | |
| | relative recall | $r_x(q)$ | 77,93% | 73,38% | 71,87% | 75,00% | 78,43% | 75,32% | 78,91% |
| | | $r_{2x}(q)$ | 96,46% | 96,46% | 96,46% | 96,46% | 95,71% | 96,31% | |
| | | $r_1(q)$ | 60,00% | 64,00% | 60,00% | 56,00% | 60,00% | 60,00% | |
| | | $r_3(q)$ | 84,00% | 84,00% | 84,00% | 84,00% | 84,00% | 84,00% | |
| | prec. | $d$ | 78,33% | 78,40% | 76,59% | 76,19% | 78,21% | 77,55% | 76,77% |
| | | $d_5$ | 76,65% | 76,69% | 74,83% | 75,10% | 76,64% | 75,99% | |
| | | Avg. abs. r(q) | 51,08% | 51,34% | 50,21% | 50,23% | 50,65% | | |
| | | Avg. rel. r(q) | 79,60% | 79,46% | 78,08% | 77,87% | 79,54% | | |
| | | Average d | 77,49% | 77,55% | 75,71% | 75,65% | 77,43% | | |

From the *Tab. 1* it is apparent, that the measures do perform very comparably.

*Tab. 2: Performance of the model of 3 last years (1095 days) publishing history within MS VSS was evaluated*

| | | | Knowledge estimation measure | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $S_{Sumqa}$ | $S_{Exposureqa}$ | $S_{DiscountSumqa}$ | $S_{qa}(t_{current})$ | $S_{qa}$ | Average | Average |
| Evaluation method | absolute recall | $r$ | 60,00% | 60,00% | 60,00% | 60,00% | 60,00% | 60,00% | 49,41% |
| | | $r_x(q)$ | 41,25% | 42,18% | 40,56% | 39,07% | 41,57% | 40,93% | |
| | | $r_{2x}(q)$ | 53,06% | 52,13% | 51,67% | 53,06% | 53,98% | 52,78% | |
| | | $r_1(q)$ | 36,11% | 30,56% | 36,11% | 36,11% | 30,56% | 33,89% | |
| | | $r_3(q)$ | 58,33% | 58,33% | 58,33% | 58,33% | 63,89% | 59,44% | |
| | relative recall | $r_x(q)$ | 72,37% | 73,38% | 71,62% | 67,07% | 73,54% | 71,60% | 73,95% |
| | | $r_{2x}(q)$ | 90,40% | 89,39% | 87,37% | 90,40% | 91,41% | 89,80% | |
| | | $r_1(q)$ | 52,00% | 44,00% | 52,00% | 52,00% | 44,00% | 48,80% | |
| | | $r_3(q)$ | 84,00% | 84,00% | 84,00% | 84,00% | 92,00% | 85,60% | |
| | prec. | $d$ | 74,43% | 74,13% | 73,60% | 73,79% | 75,04% | 74,20% | 73,21% |
| | | $d_5$ | 72,18% | 71,79% | 71,55% | 71,63% | 73,96% | 72,23% | |
| | | Avg. abs. r(q) | 47,19% | 45,80% | 46,67% | 46,64% | 47,50% | | |
| | | Avg. rel. r(q) | 74,69% | 72,69% | 73,75% | 73,37% | 75,24% | | |
| | | Average d | 73,30% | 72,96% | 72,58% | 72,71% | 74,50% | | |

Comparing to the previous table, it is apparent, that the recall and precision levels in *Tab. 2* are smaller following the fact of smaller participation of the experts contributing to MS VSS repository in shorter timeframe.

*Tab. 3: Performance of the model of last year (365 days) publishing history within MS VSS was evaluated*

| Evaluation method | | | $S_{Sumqa}$ | $S_{Exposureqa}$ | $S_{DiscountSumqa}$ | $S_{qa}(t_{current})$ | $S_{qa}$ | Average | Average |
|---|---|---|---|---|---|---|---|---|---|
| | | | Knowledge estimation measure | | | | | | |
| Evaluation method | absolute recall | $r$ | 43,56% | 43,56% | 43,56% | 43,56% | 43,56% | 43,56% | 38,77% |
| | | $r_x(q)$ | 29,86% | 29,86% | 29,86% | 31,11% | 30,65% | 30,27% | |
| | | $r_{2x}(q)$ | 38,61% | 38,61% | 39,17% | 39,17% | 38,94% | 38,90% | |
| | | $r_1(q)$ | 30,56% | 30,56% | 30,56% | 30,56% | 25,00% | 29,44% | |
| | | $r_3(q)$ | 52,78% | 50,00% | 52,78% | 52,78% | 50,00% | 51,67% | |
| | relative recall | $r_x(q)$ | 70,90% | 70,90% | 70,90% | 73,59% | 73,14% | 71,88% | 75,46% |
| | | $r_{2x}(q)$ | 90,38% | 90,38% | 91,35% | 91,35% | 91,03% | 90,90% | |
| | | $r_1(q)$ | 52,38% | 52,38% | 52,38% | 52,38% | 42,86% | 50,48% | |
| | | $r_3(q)$ | 90,48% | 85,71% | 90,48% | 90,48% | 85,71% | 88,57% | |
| | prec. | $d$ | 78,80% | 77,98% | 79,10% | 78,60% | 77,93% | 78,48% | 78,14% |
| | | $d_5$ | 77,96% | 77,45% | 78,13% | 78,02% | 77,41% | 77,79% | |
| | | Avg. abs. r(q) | 37,95% | 37,26% | 38,09% | 38,40% | 36,15% | | |
| | | Avg. rel. r(q) | 76,03% | 74,84% | 76,28% | 76,95% | 73,18% | | |
| | | Average d | 78,38% | 77,72% | 78,62% | 78,31% | 77,67% | | |

The absolute recall values in history evaluation continue their fall thanks to the shrink of analysed history timeframe to only current 1 year of publication activity, but the relative recall values as well as the precision values have risen in comparison to 3 years publication activity. Graphical expression of the average relative recall values throughout all the types of analysed document repositories is within *Fig. 4*:

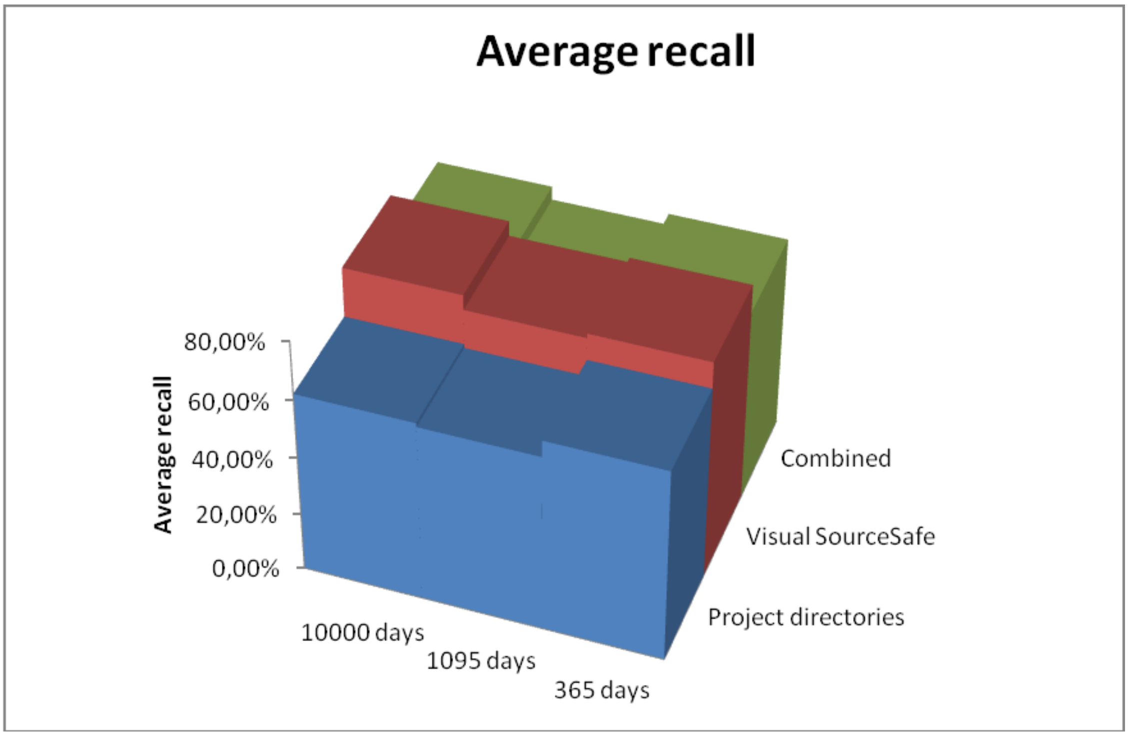


*Fig. 4: Average relative recall values in analysed document history focusing timeframes*

It allows us to formulate the hypothesis, that two effects are taking part in current knowledge assessment according to relative recall metrics; I would call them coverage and topicality effect:

- The coverage effect – In case the long term publication history is evaluated, higher recall and precision levels are achieved by having data about author's whole publication activity. The results of overall long-term knowledge become precise. The coverage effect takes place in environments, where authors have stable areas of interest and work. The more stable projects with stable stuff, the more precise results in long term knowledge assessment, which is in this case as well the precision of current knowledge level assessment. In case, where knowers have changed their projects and interests, this effect leads to misclassifications.
- But that is not the full explanation of the graph. In short term, the relative reliability of knowledge explanation grows again. The explanation is probably the topicality effect – in short term evaluation the publication activity of current period reveals current knowers very well. Although on the other side the classification of long term knowers is much weaker.

In this research both of effects took place, because of relatively stable areas of interest of people, whose knowledge was assessed, but instable enough for topicality effect to take place in short term. In case of whole history (10000 days) period of research the coverage effect causes good results, In case of one year period (365 days) the topicality effects takes over.

## 7. Conclusions

Knowledge mapping model described in this article is able to fulfil the requirements described in the section 2 of this article, i.e. mapping the knowledge of a company in an unobtrusive manner by assessing the knowledge reflected within documents of particular authors. The areas of knowledge don't have to be predefined. The model seems to be suitable for middle or large scaled organizations and organizations, whose geographical distribution is a knowledge exploration barrier.

The model does the mapping in a viable way. It allows to identify the participating current experts in the set of same size as mentioned in the survey ($r_x(q)$), with reliability ranging from 67% to 78% (depending on the measure, timeframe on history focussing) in case of all experts do publish in the document source (relative measuring on MS VSS). The performance of various measures (even not analysing the history) seems to be very comparable. The $S_{qa}$ measure seems to perform at least equally or usually outperform other measures, when longer term (3 and more years) history within a document repository (like MS VSS) is analysed and should be therefore recommended for this type of repositories. The usage of score discounting $S_{DiscountSumqa}$ does not bring any special effect, even when used on the whole document history timeframe. The usage of history focusing seems to be reliable and brings attention to further research in hypothesis of coverage and topicality effect.

The selection of document source and history coverage of the document source is predisposed to be the main factor influencing the performance of the model. The impact of history coverage timeframe on the results is apparent. The impact of document source selection was described in (Nožička, 2012). Both impacts open a space for exploration on other frequently used text sources like e-mail (when the privacy issue is solved), blogs, or wikis as the source of the model. The level of expert predictability in case all the users use the document repository is very high.

Because of the extent of this article, some of the aspects of the research performed have been omitted as: The analysis of impact of selected document source on model's reliability. Another one is the reliability of various knowledge estimation measures. Those are discussed within (Nožička, 2012). As well analysis of impact of search engine scoring method goes out of scope of this article, though it has been done and is prepared for publishing in my doctoral thesis.

Among the problems still open requiring further research is enhancing and précising the parameters of the measures to their optimal values (currently the values were determined experimentally), developing the framework for implementing the model on various types of document repositories and various types of data (semi structured or structured texts as e.g. programming source codes).

## 8. References

Busch, P. A., Richards, D. & Dampney, C., 2001: *Visual Mapping of Articulable Tacit Knowledge.* Darlinghurst: Australian Computer Society.

Jambrich, M., 2005: *Získavanie znalostí a lokalizácia expertov z neštruktúrovaných a semi-štruktúrovaných dát v kontexte manažmentu znalostí.* Praha: VŠE Praha.

MITRE, 2008: *Using Knowledge: Advances in Expertise Location and Social Networking, A Case study,* místo neznámé: APQC.

Mockus, A. & Herbsleb, J. D., 2002: Expertise Browser: A Quantitative Approach to Identifying Expertise. *International Conference on Software Engineering*, pp. 503--512.

Nonaka, I. & Takeuchi, H., 1995: *The knowledge creating company.* New York, Oxford: Oxford university press.

Nožička, J., 2003:. Mapping knowledge within an organization: KM from different point of view. *Systémová integrace,* 8(1)

Nožička, J., 2012: *The Unobtrusive way of Organisational Knowledge Mapping.* Cartagena, Academic Publishing, pp. 848 - 858.

Seid, D. Y. & Kobsa, A., 2003: Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce,* I(13), pp. 1-24.

Smith, R. G. & Farquhar, A., 2000: The Road Ahead for Knowledge Management. *AI Magazine*, 21(1), pp. 17-40.

Sternberg, R. J., 2002: *Kognitivní psychologie.* Praha: Portál.

Vivacqua, A. S., 1999: Agents for Expertise Location. *AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace*, pp. 9--13.

Yu, B., Venkatraman, M. & Singh, M. P., 1999: A Multiagent Referral System for Expertise Location. *Proceedings of the AAAI Workshop on Intelligent Information Systems*, Orlando, Florida, 18--22 July 1999, pp. 66-69.

**JEL Classification    C80, D80**

## This article should be cited as:

Nozicka, J., 2012: Current expertise location by exploiting the dynamics of knowledge. *Journal of Systems Integration* 3 (4), pp. 40 - 50. [Online] Available at: http://www.si-journal.org.
ISSN: 1804-2724