

Recommending the Right Knowledge

Cheng-Lun Li
School of Information, University of Michigan

Introduction:

Knowledge Management is crucial for most organizations. Companies maintain their core capacities and advantages through effectively utilizing their knowledge assets. However, it is hard to create a good knowledge sharing environment. Especially in large-sized companies, the sources of knowledge are usually distributed and have low visibility for the whole organization. Knowledge Management Systems (KMS) were developed as solutions to the problem. Yet, many KMS nowadays still serve as only data repositories. The data processing techniques of these systems are not sophisticated enough to help user filter the irrelevant information resources and recommend those which match users' expectations.

I propose a recommender system as a supplement to the existing current KMSs. The recommender system improves the KMS in two ways: 1) It helps the KMS to push the right information to the right people at the right time. To achieve the goal, the recommender system presents a ranked search results based on the relevancy and the quality of the document. 2) It identify the experts with different domain knowledge and connect them to other employees they can help. A contextual expertise level is provided based on the queries of the users. .

Popular Knowledge Management Systems:

There are several popular KMSs used in organizations. Lotus Notes is one of the earliest KMS, which serve as a data repository and messaging system. Users can create, edit, and share content together. Recently, Microsoft released MS SharePoint, which is a powerful KMS that incorporate both resources of data and expertise for users to search. However, to use the full functionalities, the company needs to build SharePoint server first.

Why we need a recommendation system for KMS:

A good recommender system promotes the right information to the right user at the right time. It is particularly useful when the information is too much that users need a filtering technique to prevent information overload. Unlike the keyword-based search, the recommender system considers both preference and relevance.

System Overview:

To make the system realistic, I decided to limit the target company to a consulting firm, ExpertWork. ExpertWork is a technology-based consulting firm that provides all kinds of IT solutions to the customer. Their service includes recommending options of IT systems, implementing systems, and providing them legal supports. Their major competitors include IBM consulting, and Accenture.

Characteristics of the items:

The documents as the items:

The types of documents can vary a lot in terms of its purpose, time produced in the project cycle, or simply the format set by the system administrator. Information carried by these documents also varies. When users enter queries, they also have different purposes. Some might be looking for a confirmation of technical process, while some might be looking for best practices. The strategies of recommending these documents reflect users' expectations. When users need definitions of terms, or description of certain technology, they expect the result to have high similarity to the search term. When they are looking for best practices or previous experience, the system needs to find out the relevant projects to recommend. In order to provide appropriate items, meta-data of the documents is needed to represent the property of theirs. While searching, users need to indicate their desired type of items.

The "age" of the documents also influences how they will be recommended. Both the accuracy and the relevancy might change as time pass by. The best practices five years ago might not be feasible anymore. Moreover, the improvement of the technology makes it harder to update the technology. The system should be promoting not only the most relevant but also most recent information.

The people as items:

The purpose of recommending people as items is to facilitate knowledge sharing through communication channels. Some of the problems cannot be solved through literature review. Therefore, the most efficient way to solve the problem is by asking domain experts directly.

The members of ExpertWork have diverse backgrounds and expertise. The meta-data of them can be their education, work experience, project involved before, and history of documents they created. These records can be used to represent employees' expertise. The system should have a metric to categorize the users and provide recommendation.

Users of interests:

User of KMS can be anyone within the organization. They may have different purposes to use it. Project managers use it to identify the suitable team members for different projects. Team members use it to search and retrieve useful information and also knowledgeable employees who can help them solve the problems. Depending on the complexity of the system, the users can be benefit in various ways.

The recommendation:

Two types of recommendations will be provided: documents and people. First, as mentioned above, the document recommendations can be either specific (explanations to certain technology, or terms), or abstract (the best practices or experience to similar projects accumulated in the past). The documents will be recommended with the meta-data, such as title, type of document, create time, and teaser of context of where the keywords are.

Second, the expert recommendation is done based on the expertise level to the topics of interest of the users'. The people in higher rank are expected to be more knowledgeable to the topic. The

recommendations presented will be the pointers the other users' profile, which includes the name, business title, division in the company, and built in communication channels (messaging system, chat, etc).

System design:

This section includes the detailed design specification of the recommender system. The

Overview:

The recommender system will take various information from the meta-data, explicit/ implicit usage data. Please see table 1 for detail.

The input:

	Documents	People
Meta-data	<ul style="list-style-type: none"> Title Create time Content body Tagging 	<ul style="list-style-type: none"> Business title Tagging Previous projects Previously created documents
Explicit rating	<ul style="list-style-type: none"> Ratings to the document (5 point scale) 	<ul style="list-style-type: none"> Rating of expertise level (5 point scale)
Implicit rating	<ul style="list-style-type: none"> Reading time 	<ul style="list-style-type: none"> Document download rate

Data processing:

Different algorithms will be used to generate the recommendation. As mentioned above, users may have different expectations on the same or similar search term. Therefore, the criteria of ranking documents and people are also different. This section explains the recommendations of algorithms for different part of the output.

Content-based recommendations:

Content-based filtering had been heavily used in information systems. The purpose of content-based filtering is to find out similar items, based on the search queries. In knowledge management systems, the information available for content analysis can be the content body, such as text within the documents, or the text in user profiles. The results of the content analysis are usually used to find out the relevance of the document to the query.

The user generated meta-data is also very helpful information for categorizing the documents. In our case, the system takes the tagging history of each item and either sorts them into categories, such as "XYZ software installation", or match the tag with the search terms. The information architecture is built through two types of taxonomy systems: 1) hierarchy and control vocabulary and 2) free tagging.

The controlled hierarchy and vocabulary usually represents the work process, or organizational structure. Users of the documents in the KMS are tailored to use pre-set tags and categories. The classifying is achieved through collaborative tagging.

The free tagging feature in the KMS is provided to annotate the document. The users can create new tags if they believe the documents have new properties. Therefore, new meanings can evolve as more people tag. To lower the cost in tagging, the system should provide recommended tags, based on popularity, for users to use.

The data processing techniques and the output will be discussed in the following sections.

Collaborative filtering recommendations:

Collaborative filtering in our KMS indicates the usefulness of the content. Readers of each document can rate the usefulness of the content. The concept of usefulness is complicated and can be interpreted from different approaches. For example, subjective opinions of “how relevant the document is?”, “how well the documents were written”, to “how good the solutions are to the problems?”

Considering only the accuracy, inputs from all types of evaluation scheme should be collected from the users. However, the request of too many rating activities will dramatically increase the efforts from the users. This will impede users from rating. Or, they might put random ratings in order to save time. While the required commitment of users is limited, a better way of collecting rating can be targeting the users who asked similar questions. For example, give both the overall rating and rating that came only from those who asked “XYZ installation”. In that way, the result contains both the relevance and quality. Along with the numerous ratings, the rating menu will also require users to assign one property to the documents. Properties for selection include “insightful”, “informative”, etc. The options for properties at Slashdot (Slashdot.org) are good references.

At last, it is proven that the length of reading time of the documents is highly correlated to the interest level of documents. In the organizational setting, we consider higher interested level is also highly correlated to the usefulness of the documents. Therefore, the implicit rating should also contribute to the rating of usefulness

Expertise level recommendation:

The goal of users of the people recommending feature is to identify expertise with certain skill set or experiences. They want find people with domain knowledge on the topics they are asking. Therefore, the recommending criteria should include both the “types of expertise” and the “levels of expertise”.

Area of expertise: Content based filtering techniques can be used to identify the area of expertise. Similarly, the system can use the content body of users’ profiles and other users’ tagging history to classify the area of expertise.

Levels of expertise: when making design decisions on the expertise level, we should understand the instruments that constitute the levels. Consider the situation of two users giving rating to the same user profile. we should give the person with higher expertise level more weight for their

rating, since it is intuitive to believe that people will ask questions to people they consider more knowledgeable. The measurement should include the concept of relative expertise level.

The output:

The type of output can be diverse depend on the input, the processing method, and the users' expectation. In this section, I will explain the outputs and their presentations to the users. The information that will be provided to the users includes:

Document recommendations:

The relevancy of the documents: the rank, based on the relevancy, of the documents in the results indicates how relevant the documents are to the search terms. The ranked recommendation is the first filter that selects the potentially useful documents for the users. The information that the system used to calculate the relevancy are the content body and the tags assigned by the users. (See process A for the calculation)

The rating of the documents from other people: The ratings of the document indicate the usefulness of the documents. As mentioned above, the system will only require the users to give one numerous rating and assign one property. Therefore, the score of the rating will only be supplement to the results. Users can set threshold of the relevancy and sort the average ratings on top of the remaining results. (See process B for the calculation)

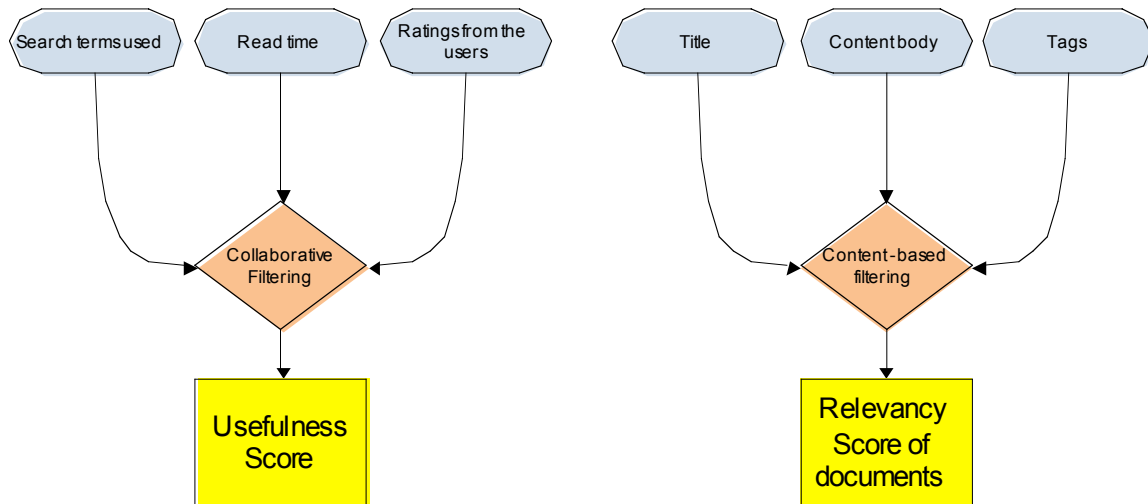


Figure 1:input/output of documentations

Expert recommendation:

The relevancy of the expertise: The first filter of the expertise searching is also the relevancy. Ranked results on the relevancy are provided for reference. The information needed to calculate relevancy is also the content body of the user profiles, and the tags assigned. (See process C for calculation)

The expertise level of other users: the KMS incorporate the concept of “prestigious” when calculating the expertise level. Depending on the expertise level of the raters, the weight of their

ratings are assigned differently. Generally, people get more prestigious status after getting good rating from people who have higher expertise level. (See process D for calculation)

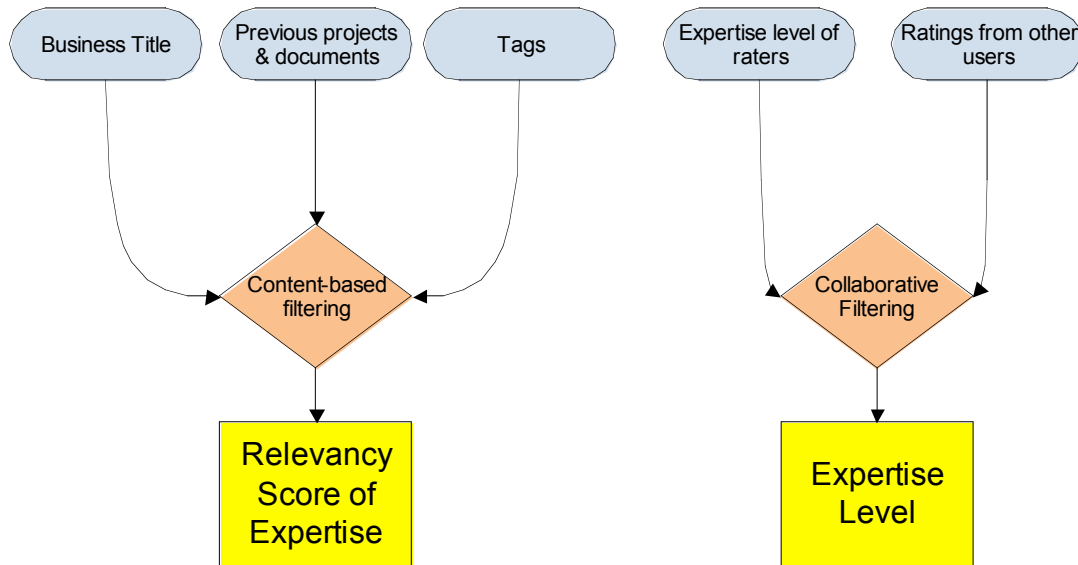


Figure 2:input/output of expertise

Design Recommendation on Algorithm:

There are four different data processing happening. Among them, three different algorithms are used.

Process A: process A uses content-based filtering. Three different inputs are used when calculating the relevancy. The algorithm used is the tf-IDF value. We compared the content body with the search terms and calculate the value of tf-IDF.

The equation of tf-IDF is:

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i$$

, where

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \text{ and}$$

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

tf is the term frequency, while IDF is the inverse document frequency.

In the equation for tf , n_{ij} is the number of occurrences of each search term in the document. The denominator is the total number of occurrences of all terms in the document. The more occurrences of search terms found in the document, the higher the tf will be.

In the equation for IDF, $|D|$ is the total number of documents exists in the KMS system. The denominator notation represents the number of documents where each of the search term appears. The value of the IDF represents the importance of term (t_i) is to all the documents.

The value of tf -IDF indicates the level of relevancy of each document to the search query. After getting the results, the value will be normalized and put together with the other tf -IDF values to the title and tagging history. The weight of different input should be in the order: title > tag history > content body. However, the details of the weighting factors need to be tested with sample data.

Process B: Since the users of KMS usually have specific expectations, which are often not related to personal interests, to the information provided from the KMSs, it is not suitable to use item-item or user-user algorithm as the filtering technique. Instead, the system gathers the ratings from other users who hold similar expectations and determine the usefulness of the document with context. For example, the system gathering the ratings to document A from the users who has similar search terms to “XYZ” and determine the contextualized average rating. The contextualized rating score is combined with the reading time and constitute the measurement of usefulness.

Process C: process C also uses tf -IDF to calculate relevancy score of expertise. The difference here is that the tags of previous documents and project description is used instead of the content body. Again, the final value is a weighted combination of the three inputs.

Process D: process D takes each rating and the expertise level of the rater to calculate the expertise level of the target user. Using the page rank algorithm, we can get the prestigious level, which is also a good indicator to the expertise level of each user. The expertise level of each user is the weighted combination of the ratings from other users. We set a threshold of rating, say 4 (knowledgeable) and create an expertise network. The edges are weighted based on the rating (4, or 5). Based on the network, we calculate the page rank value and rank the expertise level of users.

Evaluation of the recommendation:

There are several ways to evaluate the recommender system:

Accuracy:

The recommender system should have certain satisfaction rate in providing the results which the users expect. There is several ways to gather the feedback from the users. The system can prompt the question about satisfaction after users click the link to the documents. Or, it can implicitly record the users' browse patterns. The reading time of the user is a good indicator of how the documents fit their interests. For the expert recommendation, the system can also monitor users' browse pattern, but this time, log the activities history, such as copy of the text, or click rate of the link to leave messages. The monitoring data indicates the success rate of users' task. In addition, the data analyst should perform random sample testing and compare the accuracy to the golden standard, the human judgments. For example, they should calculate the correlation of normalized ratings from human and computer.

Efficiency:

The efficiency related to the computation time needed for the system to provide recommendations. Some of the calculation of ExpertWork's KMS can be done offline, such as the expertise level, while some of the calculation starts after receiving users' query. The system should run the performance test on computation time for randomly generate search terms and evaluate the efficiency. Depending on the scale of the system, it is possible for a long response time of the KMS. The system should consider getting only clusters as samples in calculation, rather than using all the information.

Limitations:

The accuracy of the expertise level:

When calculating the expertise level of users, the system only calculates the overall expertise level, regardless the topics of interests of the search terms. For now, the system relies on the relevancy level from the content-based filtering to filter the irrelevant items (user profiles). However, it is possible for the system to recommend the expert X, whose profile is relevant to the search term A, while he/she is actually expert in B.

Attackers of the system:

If the system is providing the status (expertise level), or the reward to the contribution, it is possible for users to maliciously attack the system, for example, boosting his/ her level of expertise, or simply the rating of her documents. The protection of the system is out of the scope of this paper. However, it should be studied before the system is implemented.