University of Regina

Department of Computer Sciences

Winter 2019

CS 824 - Information Retrieval

Assignment 1

Questions 1 & 2 & 3

Submitted to Dr. Yiyu Yao

By

Chao Zhang

Regina, February 3, 2019

1. DIKW hierarchy Discussion

DIKW is four initials of data, information, knowledge and wisdom which also shows a hierarchy form the lowest level data to the highest level wisdom. It is first proposed by Danny P. Wallace and now they are many versions of DIKW. Basically, it contains the structure composed of four components. In other versions, the structure may add other components or delete some original parts but generally the classical chain begins from data which is raw material to information to knowledge and finally reaches wisdom - a highly extracted treasure.

Information retrieval is a process of obtaining useful information from a query. It covers the data hierarchy and information hierarchy in DIKW model and also provides transition to knowledge hierarchy. Information retrieval system gives representation by users' needs through collected raw data which is a kind of travelling from the lowest hierarchy data to a higher hierarchy information. Our purpose in DIKW model is to reach the highest hierarchy wisdom and information retrieval system helps us upgrade the level. As we known, form the information retrieval system, we have knowledge retrieval which lets us to climb to a higher level in DIWK pyramid.

In conclusion, information retrieval system in DIWK model plays a significant role, it both transfers data hierarchy to information hierarchy and paves the trip to knowledge hierarchy.

2. Boolean Retrieval Models Discussion

Boolean retrieval model is a math based model which converts retrieval operation into math calculation according to set theory. It classifies the information retrieval into three main sets: T(index term set), D(document set) and Q(query set). Index term set contains terms which are key words or stems basically used for describing documents. Document set is a set of all documents. Query set is the combination of all needs form users by Boolean expression. The whole Boolean retrieval model provides us a full retrieval system. We can use query set to search document set by index term set to find the answer. In practical operation, we first obtain elements in query set and use each element to obtain different document sets accordingly. Then we use logic operation to process each document sets to get the result.

Boolean retrieval model is so classical and even adopted by people nowadays due to it has some great advantages. The model has clean format and easy to implement making retrieval effective and efficient. However, it is more like data retrieval than information retrieval and because of Boolean expression it is difficult to translate some queries.

Boolean retrieval model is a basic model and very useful in information retrieval. It both has advantages and disadvantages. For some disadvantages such as all terms in Boolean retrieval model are equally weighted, it has a upgraded version called extended Boolean model or fuzzy information retrieval model. Boolean retrieval model contributes a lot in information retrieval.

3. Report of Evaluating IR systems: Google, DBLP and Web of Science

Google, as we all known, is a famous search website for almost everything. DBLP is bibliography searching website for computer science mainly. It was established in 1993 and has grown from HTML collections to bibliography site now. Web of Science is a search website offering citation indexing service.

For similarities, they are all search websites and provide information retrieval. However, DBLP only offers computer science bibliography searching, Web of Science only offers scientific citations searching but Google can be used for searching everything. In addition, DBLP and Web of Science can only be used for searching texts, however Google provides indexing not only for text but images, music, maps, videos and so on.

DBLP and Web of Science focus on bibliography indexing so they may be lack of some media searching. When users want to find a book without name but only the cover of that book, it is perfect when those information retrieval system has cover image indexing to help people find their needs quickly. For Google which we use almost everyday, as a user, I would like developers could add a filtering function to make Google work more effectively. I find a problem when I use Google that is it always provides some unnecessary information and sometimes its ranking can not satisfy me. If Google has a filtering function, for example I would like to find the answer by searching the questions by filtering the key word "Q&A". The searching results will be mostly websites related to academic forums instead of shopping websites or company home pages. In addition, these three information retrieval systems are all lack of similarity scores function. I would like search engines can provide users with index term similarity scores to help them better know the relevance of searching results.

For Google, it can be integrated into a more general system because of its diversification. Besides retrieval function, it can add more applications. But for DBLP and Web of Science, due to the reason they are based on academic references indexing, it is difficult to be a more general system.

These information retrieval system do offer us a great convenience, it helps us better know the world. Just like the DIKW hierarchy model, IR systems help us to climb the top of pyramid to get the wisdom and this climbing process is assisted by information retrieval systems. As a graduate student, I think high level education students or researcher whose major related to computer science or math & statistics or related fields may learn information retrieval, we all witness the transition from file systems to database to information retrieval to knowledge retrieval. Information retrieval is a significant part in computer science development and an essential course for us to learn.