University of Regina

Department of Computer Science

Winter 2019

CS 824 - Information Retrieval

Assignment 5

Submitted to Dr. Yiyu Yao

By

Chao Zhang

Regina, April 10, 2019

1. In the context of relevance feedback and linear retrieval models, discuss the advantages of acceptable ranking.

Based on relevance feedback and linear retrieval models, we have query $\vec{q}_0$ given by a user and $\vec{q}_1 = \alpha\vec{q}_0 + \beta\frac{1}{|S^+|}\sum_{d\in S^+}\vec{d} - \gamma\frac{1}{|S^-|}\sum_{d\in S^-}\vec{d}$. If a term appears in a relevant document, then the term is viewed as useful which increases the weight. Inversely, if a term appears in a non-relevant document, then the term is viewed as not useful which decreases the weight. From the definition of perfect ranking, we know it indicates $\vec{d} > \vec{d}' \Leftrightarrow \vec{d}\cdot\vec{q} > \vec{d}'\cdot\vec{q} \Leftrightarrow (\vec{d}-\vec{d}')\cdot\vec{q} > 0$ which shows each preference pair gives an inequality. It also induces $\vec{d}_1 \sim \vec{d}_2 \Leftrightarrow \vec{d}_1\cdot\vec{q} = \vec{d}_2\cdot\vec{q} \Leftrightarrow (\vec{d}_1-\vec{d}_2)\cdot\vec{q} = 0$ which shows each indifference pair gives an equality. From the definition of acceptable ranking, we know $\vec{d} > \vec{d}' \Rightarrow \vec{d}\cdot\vec{q} > \vec{d}'\cdot\vec{q} \Rightarrow (\vec{d}-\vec{d}')\cdot\vec{q} > 0$. Perfect ranking requires $(\vec{d}-\vec{d}')\cdot\vec{q} > 0$ for all preference pairs and $(\vec{d}_1-\vec{d}_2)\cdot\vec{q} = 0$ for all indifference pairs. However, acceptable ranking only requires $(\vec{d}-\vec{d}')\cdot\vec{q} > 0$ for all preference pairs. Acceptable ranking is not as strict as perfecting ranking, so it is more practical and realistic in use. Because it does not need $(\vec{d}_1-\vec{d}_2)\cdot\vec{q} = 0$ for all indifference pairs as we discussed above, so acceptable ranking allows re-arrangement in equivalence classes in documents.

2. Discuss potential applications of what you learned about information retrieval for your future study/research/work/career.

Information retrieval covers a large range of knowledge. From the indexing, I have learned ranked list of words, frequencies of words and removal of high and low frequencies. It helps me better research in text mining in the future. I am going to work as a data analyst after my graduation. Learning analyzing frequencies is so significant. From system evaluation and relevance feedback, I have learned some concepts like binary model, user preference model, perfect ranking and acceptable ranking, error correction and so on. They are also more or less related to artificial intelligent and machine learning. We all know AI will be a hot topic in the future. Some concepts in information retrieval formulates the basic knowledge related to AI. In the future, it is more easier for me to get started to work in this field quickly because of having background in it. It seems like all the topics I have learned in information retrieval are necessary in search engines optimization. The class gives me a beginning of knowing the word of search engines. After more researches in information retrieval, there is a possibility in finding a job in search engine companies like Google. In addition, the magic 3 is also covered in information retrieval. The basic structure of information retrieval has 3 parts: documents, queries and matching. The computation in system evaluation has 3 cases: agree, contradict and compatible. This magic 3 shows many times in information retrieval and it also affects my daily life. Sometimes, thinking something in 3 makes it easier. For example, the price can be thought as low, median, high. The quality of product can be viewed as great, bad and ok. Expect from knowledge in information retrieval and math models, we can apply this magic of 3 into real word. Just like 3 cases in calculation of ndpm, we can divide our problems in real life into 3 situations and response to them with different methods. Learning knowledge from abstract models and applying it into practical things is the best treasure I have obtained from this class.

3. Write a two to three pages introduction to Information Retrieval.

Information retrieval is obtaining information from resources according to need. The basic structure of information retrieval can be three main parts: documents, users and matching. Documents need representations. Users provide queries. Matching combines documents and users' queries to provide retrieved information. Matching can be classified into two parts: exact matching and inexact matching. In exact matching, we have Boolean model. In inexact matching, we have vector space model and probabilistic models. Vector spacer model includes binary vector space models and weighted vector space models. For the results which are retrieved information, we have relevance feedback. It contains binary relevance feedback viewing documents as relevant or non-relevant and user preference feedback including perfect ranking and acceptable ranking. Relevance feedback as well as system evaluation, they both improve the effectiveness and efficiency of the whole information retrieval system.

The history of information retrieval originates from file system, then to databases and finally to IR system. Further more, information retrieval can reach knowledge retrieval. For each of part in the whole structure discussed above, it has many models and concepts. In the beginning, we have documents. From documents to results, we need indexing.

Indexing is a transition from external representation like web pages to internal representation such as terms. We determine the set of terms by controlled vocabulary and uncontrolled vocabulary. Then we have a basic assumption of relevant and non-relevant according to if there is a word appears in documents with its frequency determining the importance.

From indexing, we know terms and frequencies. So, a term to frequency matrix can be built which is TM matrix in short. Then we introduce Boolean model with binary document representation and Boolean query expression. In order to avoid all terms are equally weighted, we need to change TF matrix into a weighted matrix. There are many weighted methods: row-wise normalization, column-wise normalization, R*C, TF*IDF and so on.

To improve the system, we then have system evaluation. Evaluation is to find measures which tell 1. whether a system is good; 2. if one system is better than another system. There are three classical measures in binary evaluation: precision-based, recall-based and balanced. In user preference system evaluation, we then discuss user preference relation, indifference relation, perfect ranking and acceptable ranking. The measurement for user preference evaluation can be calculated by ndpm. Ndpm classifies the relationship between system and user into 3 cases: agree, contradict, compatible which has 0, 2 and 1 credits respectively. By calculating the whole sum, we can know the exact matching degree.

For inexact matching, there is vector space model of information system. In vector space model, we have different weighted formula to represent a document, cosine similarity and linear retrieval function as well as probability interpretations. At last, we introduce relevance feedback. It asks the system learn from examples and correct errors by itself to construct a bridge between results and documents.