Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
Lecture 12: Completion of relevance feedback; implicit feedback sources
3/7/06
Lecturer: Lillian Lee
Scribes: Nicholas Hamatake (nh39), Min Park (mp257)

---

1 Introduction

1.1 Overview

We have been discussing Automatic Query Expansion (AQE) and Interactive Query Expansion (IQE) as applications of *explicit relevance feedback* in the past two lectures. In particular, the last lecture spent some time on the comparative takes of IQE versus AQE; one of these included a discussion of Ian Ruthven's SIGIR paper, "Re-examining the Potential Effectiveness of Interactive Query Expansion" which received Best Paper award in 2003 ([RI]).

Much interest was expressed by students of CS 630 on Ruthven's study; so, today's lecture will start by elaborating on the details of his study, leading to some interesting points and questions regarding IQE. We then consider '*explicit relevance feedback* conclusions'*:* a summary of some of its current issues and possibilities.

One major concern with explicit RF is that it is expensive. What about *implicit RF?* We could infer relevance labels for documents based on users' behavior. Noting its obvious benefit as being less expensive to the user, three examples of information being used as 'user behavior' are presented; they are placed in the context of arriving at relevance judgments for documents. The third example is more involved and is explored in greater detail in the next lecture.

2 Relevance Feedback

2.1 A quick review

Recall that having relevance feedback means that we have relevance labels on documents for a given query. Moreover, we have been assuming that it is explicit in the sense that it is provided directly by the users and entails binary judgments for a certain set of retrieved documents. Relevance feedback information can be used in future queries both to re-weight terms and to discreetly insert terms into the queries themselves. This *query expansion* can be done either automatically (AQE), or through interactive feedback from the user for each potential term (IQE).

While AQE can be confusing to the user (especially when it 'goes wrong'), effectiveness of IQE also depends on how well users can choose expansion terms with respect to their effectiveness in the system. That is to say, what a human intuitively believes (or assumes) will be the 'correct' terms for improving queries may not actually yield better search results. Consideration of this human aspect of IQE makes it challenging to gauge the relative merits of AQE and IQE. Ruthven's study (discussed in the following Section) is nonetheless able to provide some insightful results.

2.1 More on Ruthven's '03 AQE vs. IQE study

2.1.1 The Best Potential IQE System

(Section Reference: [RI]; this section includes background information for Ruthven's study, additional to lecture discussion)

How does Ruthven study the best potential IQE performance? He considers sets of queries across three corpora (Associated Press, San Jose Mercury, and Wall Street Journal) for which an expansion *could* augment retrieval performance. Specifically, he excludes queries for which *all* or *none* of the relevant documents were within the top 25 retrieved documents. Note that *perfect relevance feedback* was assumed.

For each *q*, a list of top fifteen expansion terms was created[1] based on relevance judgments for the top 25 retrieved documents. All 32,678 possible query expansion decisions were carried out for *q*. The performance of each IQE decision was ranked by average precision, the best in this ranking being the 'best IQE decision' for each query (while 'middle IQE' and 'worst IQE' decisions are the ones ranked 16, 384th and lowest in this ranking, respectively). He considers the *percentage of the queries improved* when 'best', 'middle', or 'worst' IQE decisions are made for each query, relative to when there is no query expansion.

By considering the best IQE decisions for each query, he is considering the performance of a "perfect system" in which query expansion decisions are always the best-performing ones in the system. He also compares the performance of *all* IQE decisions to the percentage of queries improved using the best of the three AQE strategies[2], which leads to his conclusion that it may be difficult for humans to make better decisions than AQE. He notes, however, that in this consideration of all possible IQE decisions, an inherent assumption is being made that systems lack any method of selecting good combinations of terms (as a human might). He investigates how humans might *actually* rate terms presented to them in a simulated IQE setting.

2.1.2 Good with respect to performance vs. good with respect to human judgment?

For each query, the top fifteen possible expansion terms were 'manually' (automatically for each term) classified as *good*, *poor*, or *bad* in the following manner:

For each term:

- If the *average precision* with the term was **greater** than without it, then the term was considered '**good**' by the system.

Similarly,

- If the average precision with the term was *less* than without it, then the term was considered *poor*.

Note that a 'poor' term means that the term was poor with respect to retrieval performance in query expansion, but still had a high selection score since it was among the top fifteen expansion terms considered. That is, the underlying system originally presented it as a good term, but when included in query expansions, it performed worse than queries that did not include this term.

Three human subjects were given three sets of the same eight queries selected from each corpora (so 24 queries), as well as some context for each search[3] so that they could understand what the information need was for the search task. They were then presented with the fifteen possible expansion terms for each query

---

1 More than fifteen terms would have been computationally too intensive.

2 Query dependent AQE

3 They were shown the initial queries, full TREC topic description, and were also asked to read each of the relevant documents found within the top 25 retrieved documents that were used to create the list of the top 15 expansion terms. However, subjects were *not* asked to read non-relevant documents because '[Ruthven] felt it would be too great a burden on the subjects' [RI].

(so 360 terms), and asked to rate each as 'useful' or 'not useful' in terms of improving that query; 'cannot decide' was another option in their response. Ruthven was trying to simulate a 'realistic' environment by presenting users with terms that an IQE system might actually present them. The tabulated results from Ruthven's paper are reproduced in figures 1 and 2.

	Subject 1	Subject S2	Subject S3
AP	73%	60%	63%
SJM	50%	40%	42%
WSJ	62%	32%	45%

Figure 1: Percentage of good expansion terms detected by subjects [RI]

	Subject 1	Subject S2	Subject S3
AP	54%	36%	43%
SJM	39%	26%	35%
WSJ	38%	45%	39%

Figure 2: Percentage of poor expansion terms classified as good by subjects [RI]

The results suggest that what a human considers to be 'good' does not imply that the term is 'good' with respect to performance in the system. To summarize these results, 32% - 73% of good terms were detected correctly, and 26% - 54% of poor terms were detected *incorrectly* with variability across users and corpora. That is, the subjects of this study seemed to have a hard time picking terms that are *actually* useful for query expansion.

With regard to this small experiment and its outcome, we could ask several questions; here are three:

1. *What if the users chose everything?*

   In other words, what if the users (in this study) were not judging each term carefully and selecting most of them as useful, whether they seemed good, neutral or bad? This certainly seems to be a possibility given that so many terms—24% to 54%—that perform poorly were selected as good according to the users.

2. *How good is the resulting query?*

   Were the human IQE decisions tested in the system? Data produced on average performance of these human queries would be interesting and might be informative.

   Some conjectures were made in class as to the absence of this information. It may have been the case that the small size of the data would make these results difficult to interpret, or perhaps the author believed that the users' judgment of terms was good enough.

3. *On what basis are terms being chosen? Who are the users?*

   While there was low correlation between human and system ratings of a term as good/poor/neutral, within the small data set, there was a large correlation between what the users picked and term frequency in the corpus!

   Users had a tendency to rate common terms as useful. Ruthven notes that Subject 2 (of the three) did the opposite with terms in the Wall Street Journal ([RI])—also notice that the lower bound on the range of good terms detected, 32%, is due to Subject 2's term ratings for the Wall Street Journal.

   While there is no description given about the subjects in Ruthven's paper, it may also be interesting and/or informative to consider who the subjects were.

The debate on the effectiveness of IQE given relevance feedback information raises many interesting questions and points. More are considered in the enclosed 'questions' section of this lecture guide.

2.2 Explicit RF Conclusions

2.2.1 Short term utility: "FOR YOU, RIGHT NOW, THIS QUERY"

- Explicit relevance feedback is query-specific and does not seem transferable
- Explicit relevance feedback is also user-specific:
  For two queries that the system considers the same, users can have different relevance judgments on the set of retrieved documents (Teevan, Dumais, Horvitz '05). Example: "Key documents" could be considered different from "Important documents" to some user.

In these ways, explicit relevance feedback does not seem easy to generalize, although counterexamples exists, such as queries with (mis)-spellings which are often unrelated to a user's intended meaning of a query term.

2.2.2 Long term utility?

- RF is expensive, particularly for the user.
- Even with RF, it is difficult to have better retrieval (to 'get it right')

Nonetheless, a plausible long-term benefit of gathering RF data is in the chance that it could accumulate enough data to be able to create 'user profiles', in which case one might even imagine having enough information to categorize users into certain *user types* based on these profiles.

An important note to make is that RF is expensive. This succinct point is often used as a strong argument against it. Could there be other (less expensive) ways of obtaining relevance labels for documents given a particular query?

3 Implicit Relevance Feedback

3.1 What is it?

Up until now, we have assumed that relevance feedback information is given explicitly by the user ('Yes' or 'No'). In the *implicit relevance feedback setting,* similar information is *inferred* from information relating to the user's *behavior,* and not by direct engagement with the user. The following sections consider examples of information drawn from users' behavior that could be exploited to infer what the user's relevance labels (or weights) for some documents might be.

An obvious benefit of implicit RF is in that it is less expensive to the user. Still, it involves an interactive process with the user. This is ok, however, since query-log studies show that users tend to re-formulate queries (not surprisingly), and this is a form of interaction with the system that also provides some information on relevance (discussed in greater detail in section 3.3.2 and 3.4).

3.2 A Canonical Example: Reading Time

What is explored in this example is the question of whether or not there is a correlation between reading time (≈ display time), and relevance of a document. Two arguments are presented. The first is based on a study that indicates that there is a correlation, while the other is a contrasting study that claims that there is no correlation. If there is correlation, it could be used for implicit RF.

3.2.1 Morita and Shinora's Experiment ('94): Reading Time for Newsgroup Articles

The subjects in this study were eight users who were asked to read all of the documents from newsgroups they were subscribed to for a period of six weeks. This results in data with information on about 8,000 total documents—about 24 documents for each person read each day.

The idea of this study was to understand the effect of interest on reading time (measured by display time of a document), and use it as a sort of proxy for the correlation between (apriori?) relevance judgments of a document and reading time.

Controlling factors considered for the study included document length and complexity, and length of queue. Another factor that may have been used—but we are not certain was used—is user reading speed. They were implemented in hopes of being able obtain a more isolated measure of the effect of interest on reading time; that is, *after* controlling for some known factors that could affect display time. For example, if a document is long (and/or complex), or if their queue is very long, the user may not read a document as thoroughly as they might if the queue and/or the documents were shorter.

The study concluded that there is correlation between reading time and relevance.

3.2.2 Kelly and Belkin's Refutation ('04)

Kelly and Belkin claimed that reading time and interest level are not correlated. In their study, they argue that *reading time* is task-dependent, and not based on interestingness, and that it also depends on the user. The following example was presented in lecture (note that it is not from the paper).

Consider a user who is searching for the height of Mount Everest, and finds this information relatively quickly in a very long document. Then this document was relevant; however, the user may have only spent moments viewing this document for the necessary information, and may not be interested in the document overall. A more complex information need, however, may require a very thorough perusal of a document before a user can even judge whether or not the document meets their information need.

3.3 Example: "Stuff I've Seen" / Google Desktop

3.3.1 The Desktop as the Source of Information: Teevan, Dumais, Horvitz, '05

"Stuff I've Seen" (similarly with Google Desktop), indexes everything on a user's desktop including:

- all information that the user creates and views on their Desktop
- webpages viewed
- e-mails
- calendar items

Given this great deal of information on the user, various portions of it could be viewed as relevant. One could use all of the information, only the most recently-accessed items (the study considered a one-month window), or only a certain type of item (the study considered using only webpages). These three options were compared and the authors found that, somewhat surprisingly, using *all* items yielded the best performance. (A note on implementation: the authors used the RSJ model with weight updates and automatic query expansion to incorporate the implicit relevance information.)

Note that a similar pilot study was done in 1999 by Budzik and Hammond, where the intent was the use the text *currently being typed* by the user as implicit feedback.

3.4 Query History and Clickthrough Data – the "poor man's version" of Stuff I've Seen

What if such extensive personal data is not available? In particular, what "documents" can implicitly be considered as relevant at the beginning of the iterative interaction with the user? As with explicit RF, we can always use the query itself as a relevant "document."

Better still, we can use the query history (or a cross-section of it) as relevant documents. Given a user's complete query log, it is important to distinguish which queries represent attempts to satisfy the user's current information need, from previous and different ones. Different researchers employ different methods; one study (involving Thorsten Joachims) observed positive results using only queries issued within the last 30 minutes.

In addition to query data, it also seems natural to consider a user's decision to click on a link and view the resulting document as an implicit endorsement of that document. So-called 'clickthrough data' can be used in much the same way as a query log. The most straight-forward approach is to consider as relevant each document that a user views after being presented with the results of his search. Alternatively, one could only treat the corresponding document summaries as relevant mini-documents, since only the summary was available to the user at the time he made the decision to follow the link. In practice, both of these methods are employed and no conclusive study shows that one should always be preferred over the other.

3.5 Incorporating implicit feedback data

Once we have implicit feedback data, what do we do with it? Obviously, one could treat it just like explicit RF using techniques discussed in last week's lectures.

One alternative approach to this was proposed by Shen, Tan, and Zhai in 2005. They proposed a method of incorporating implicit RF into their original language model framework. The basic idea is to score a document by the distance between the language model that produced that document, and a new language model P_k^*, representing the estimated information need after k iterations with the user. In each round, query history data and clickthrough data is used to update this model of estimated information need. The details of the study will be given next time…

4 References

[RI] Ruthven, I. Re-examining the potential effectiveness of interactive query expansion. SIGIR, pp. 213-220 (2003)

5 Questions

It is difficult to conceive basic "finger exercises" for this material, since no new methods of actually incorporating RF into various IR models are covered in detail. Our exercises, therefore, are more conceptual in nature.

1. Ruthven's experiment on three human subjects.

   Motivation This experiment was realistic in the sense that the subjects were presented with terms that they would have been presented if they were searching in a system using IQE. Nonetheless, it was artificial in the sense that these searchers did not 'belong' to the subject. The subjects had to first become familiar with a stranger's query and that stranger's context for the query—someone else's information need. Moreover, the study does not indicate that the expanded queries as suggested by each subjects' ratings of expansion terms were actually tested.

   Rating the terms well is not the end goal of this 'stranger', the original issuer of the query, while it is that of the subjects in Ruthven's experiment. In an operational setting, this stranger issues a query for a reason, and it seems plausible that he or she would at least have some incentive to *try* to make IQE decisions that he or she truly believes will help the performance of the query—that is, return more relevant documents.

   The Question Suppose a very similar experiment as in Ruthven's paper could be carried out. Suggest (an inexpensive) way of modifying the experiment such as the subjects might perform slightly more 'realistically'.

2. Query History

   Given a list of all queries issued by a user during a given session, it behooves us to find a good way of determining which groups of queries represent a refined attempt at satisfying the same information need, and which queries represent completely different information needs. Techniques for doing this could range from the very simple (e.g., assuming all queries issued within the last 30 minutes are related) to the more complex (e.g., actually clustering queries according to some distance metric). Propose some techniques of solving this problem, and give some advantages and disadvantages of each.

6 Answers

1. Ruthven's experiment on three human subjects.

Our statement of the question may have been somewhat vague. Even so, it is intended as somewhat of a small thought exercise.

The idea here is to modify the subject's goal in a way that their behavior would more closely mirror that of the original issuer of the query (the 'stranger').

One possibility arises from a suspicion the subjects in Ruthven's study were being careless in their decisions. Suppose that it was indeed due to a lack of incentive on behalf of the subjects. While we do not know how the subjects were compensated (how much were they paid to participate in the study?), if they were, maybe their monetary compensation could depend on how well their query expansions perform!

2. Query log segmentation:

a) One could represent each query as a vector and examine the angle between each new query and the previous query issued. When the angle rises above a threshold, assume the user has moved on to a new information need. Advantages: computationally easy; doesn't require storing the entire query log. Disadvantages: false positives (detecting a shift in information need when one didn't occur) might be common due to synonymy—may need to project into a lower-dimensional subspace (or apply some other compensatory method), which would require more computation and tuning.

b) Incorporate query length. It seems that when a query produces less-desirable results, a very common refinement is to add words to the query with the intent of making it more descriptive. When a query's length increases gradually, one could assume that it is a refinement of previous queries. When a very short query is issued after a string of long queries, one could assume that the information need has changed. Advantages: simple to implement and computationally cheap; disadvantages: false negatives would occur when a user decides that adding words is ineffective for satisfying their current need and reformulates a short query with different words.

c) Perform some form of clustering. Maintain a set of clusters representing different information needs. Each time a new query is issued, add it to the "closest" cluster or create a new one if no clusters are reasonably close. One could use VSM-style cosines as the distance metric, or any other reasonable metric. Advantages: sophisticated approach, probably better performance than simpler methods; disadvantages: requires more computation, more implementation decisions, and more data that needs to be stored on a rolling basis.

Of course, there are many other conceivable options. These are meant only as a few possible solutions.