Finding Literary Themes With Relevance Feedback

Aditi Muralidharan
Computer Science Division
University of California,
Berkeley
aditi@cs.berkeley.edu

Marti A. Hearst
School of Information
University of California,
Berkeley
hearst@ischool.berkeley.edu

ABSTRACT

A common task in text analysis is find conceptually-linked passages of text such as examples of themes, imagery, and descriptions of events. In the past, researchers looking to find such passages have had to rely on searching for sets of keywords. However, themes, descriptions, and imagery may surface with many different phrasings, making retrieval based on keyword search difficult.

We investigated the application of relevance feedback to this problem. First, we implemented a relevance feedback system for sentence-length text. Then, we evaluated the system's ability to support gathering examples of themes in the works of Shakespeare. Participants with at least undergraduate backgrounds in English language or literature used either our system $(N = 11)$ or keyword search $(N = 12)$ to retrieve examples of a theme chosen by a professional Shakespeare scholar. Their examples were judged on relevance by our expert collaborator. Our results suggest that relevance feedback is effective. On average, participants with relevance feedback gathered more sentences, and more relevant sentences, with fewer searches than participants with keyword search. However, a larger study is needed to establish statistical significance.

Author Keywords
Information retrieval; relevance feedback; text analysis;

ACM Classification Keywords
H.3.3. Information Search and Retrieval: Relevance Feedback

INTRODUCTION

Conceptually-linked passages of text are central to text analysis. For example, journalists and intelligence analysts might seek quotes, excerpts, and mentions of events. Legal scholars may want to find evidence of particular treatments given to issues. In the humanities, literature scholars may search for examples of themes, imagery, particular kinds of word usage, occurrences of tropes or stereotypes, and other patterns of language use [1].

In current information retrieval systems, such passages of text are retrieved through keyword search. In system such as Google Books, the searcher types in a query, and receives a ranked list of matching excerpts from various books. No matter how sophisticated the ranking algorithm, such systems rely on searchers to produce search queries that accurately express their information needs.

For passages with a common conceptual link, generating representative keyword queries is problematic. The reason is familiar: the vocabulary problem [2]. This refers to the phenomenon that the same concept can often be expressed with many different words, and that different people are unlikely to choose the same way of describing the same thing. For example, take the Shakespearean theme that seeing an event first-hand is more credible than hearing about it from another person (discussed by Robert B. Heilman in [5] pp.58–64). This theme surfaces in two very different sentences, which have no words in common:

> I would not take this from report; it is,
> And my heart breaks at it.
> (King Lear Act 4, Scene 6 Lines 136–137)
>
> Then have you lost a sight, which was to be seen, cannot be spoken of.
> (The Winter's Tale, Act 5 Scene 2 Line 47)

Thus, relying on a set of search terms generated by a single person could lead to missing examples and non-representative results. This is a problem if the scholars seeking these passages are doing so as part of a sensemaking process [6]. Incomplete or non-representative examples compromise the integrity of arguments made using them.

A better approach would rely less on the query formulation process. Relevance feedback [7, 9] offers such an approach. In a relevance feedback system, the searcher can give the system feedback on the relevance of the results. The system uses the feedback to adapt its model of the user's information need, and return results that are "more like" the ones marked relevant. So, instead of having to formulate and reformulate queries in search of relevant results, the searcher has the option to mark relevant and irrelevant ones, leaving the re-formulation to the system. This approach takes advantage of *recognition over recall*. This is the psychological finding that it is usually easier for a person to recognize something by looking for it than it is to think up how to describe that thing [4].

Our goal was to investigate whether relevance feedback is an effective solution to this problem. Due to our existing collaborations in the field of digital humanities, we focused on the specific problem of finding examples of literary themes. Specifically, are searchers equipped with relevance feedback more effective at finding examples of themes than searchers without?

Proceedings of HCIR 2012, Cambridge MA, Oct 4-5, 2012

To answer our question, we implemented a simple relevance feedback system for sentence-length text and compared it with a keyword-search-only system in a user study. In the study, participants were asked to find examples of a pre-selected theme by searching over the complete works of Shakespeare, and given five minutes in which to do so. The theme was described in words, and they were also given two example sentences that illustrated the theme. Participants used either the relevance feedback system or the keyword-search-only system to complete the task.

After all the users had participated, the sentences they chose were rated as ”relevant” or ”not relevant” by a professional Shakespeare scholar. We then compared participants' scores across the two interfaces to determine whether relevance feedback produced improvements over keyword search.

SYSTEM DESCRIPTION

We implemented the Rocchio algorithm for relevance feedback [7], as described by Salton and Buckley in [9].The only difference was that our “documents” were in fact sentences. Our system used the vector space model of information retrieval. Each search query was translated into a feature vector Q, and sentences retrieved by ranking their feature vectors in order of their inner products with the query vector.

We used a bag of words feature vector to represent sentences. A sentence s would be translated into feature-space vector S:

$$S = (w_1, w_2, \ldots, w_n), \tag{1}$$

where w_i is the weight of word i in the sentence. Word weights were computed using tf-idf [8] with the “documents” being sentences:

$$w_i = TF(i,s) \times IDF(i), \tag{2}$$

where

$$TF(i,s) = \frac{\text{\# times word } i \text{ appears in } s}{\text{\# words in } s} \tag{3}$$

and

$$IDF(i) = \log\left(\frac{\text{total \# sentences}}{\text{\# sentences in which word } i \text{ appears}}\right). \tag{4}$$

Since we had access to part-of-speech tagged text, we were able distinguished between some words that shared the same lexical form. For example, compare the sentences, “I lost my rings.” and “The bell rings.” Because of part of speech tagging, the feature corresponding to “rings” in the first sentence (a noun) would be different in our system from the feature corresponding to “rings” in the second sentence (a verb).

The query vector Q was:

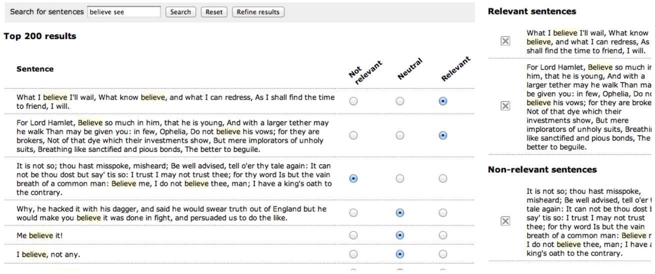
$$Q = (q_1, q_2, \ldots, q_n) \tag{5}$$

Where $q_i = 1/(\text{\# non-zero features})$ if word i was in the query, or 0 if word i was not in the query. Since the query was not part-of-speech tagged, the vector contained equal weights for all part-of-speech variants of the query words.

When sentences were marked relevant or not relevant, the query vector Q was adjusted toward the relevant sentences, and away from the irrelevant sentences, according to the following weights:

$$Q' = Q + \frac{1}{n_1} \sum_{\text{relevant}} \frac{S_i}{|S_i|} - \frac{1}{n_2} \sum_{\text{not relevant}} \frac{S_i}{|S_i|} \tag{6}$$

where n_1 was the number of relevant sentences and n_2 was the number of non-relevant sentences. Q' was then used to retrieve more sentences.



Top 200 results

Sentence	Not relevant	Neutral	Relevant
What I believe I'll wail, What know believe, and what I can redress, As I shall find the time to friend, I will.	○	○	●
For Lord Hamlet, Believe so much in him, that he is young, And with a larger tether may he walk Than may be given you: in few, Ophelia, Do not believe his vows; for they are brokers, Not of that dye which their investments show, But mere implorators of unholy suits, Breathing like sanctified and pious bonds, The better to beguile.	○	○	●
It is not so; thou hast misspoke, misheard; Be well advised, tell o'er thy tale again: It can not be thou dost but say' tis so: I trust I may not trust thee; for thy word Is but the vain breath of a common man: Believe me, I do not believe thee, man; I have a king's oath to the contrary.	●	○	○
Why, he hacked it with his dagger, and said he would swear truth out of England but he would make you believe it was done in fight, and persuaded us to do the like.	○	●	○
Me believe it!	○	●	○
I believe, not any.	○	●	○

Relevant sentences

☒ What I believe I'll wail, What know believe, and what I can redress, As shall find the time to friend, I will.

☒ For Lord Hamlet, Believe so much ir him, that he is young, And with a larger tether may he walk Than ma be given you: in few, Ophelia, Do nc believe his vows; for they are broke Not of that dye which their investments show, But mere implorators of unholy suits, Breathi like sanctified and pious bonds, The better to beguile.

Non-relevant sentences

☒ It is not so; thou hast misspoke, misheard; Be well advised, tell o'er tale again: It can not be thou dost say' tis so: I trust I may not trust thee; for thy word Is but the vain breath of a common man: Believe I do not believe thee, man; I have king's oath to the contrary.

Figure 1. The relevance-feedback user interface. Participants could mark results of a search query either relevant or not relevant (or leave them as neutral). The marked sentences appeared on the right side of the screen. If a non-zero number of sentences was marked, the “Refine Results” button would appear. Participants could click it to refine their results. If not, they could re-formulate their search query and Search with the Search button

USER INTERFACE

We implemented a minimal user interface for our relevance feedback system, shown in Figure 1. The system could be initialized either with a set of examples or a search query. Users were given the following controls: a search box, a “Search” button, a “Reset” button, and a “Refine results” button.

Searching or Refining produced a list of search results in a table. For each result, participants could mark one of three relevance-feedback radio buttons: relevant or not relevant (or leave the default choice, neutral). If a sentence was marked anything other than neutral, it would appear in a list on the right side of the screen. Participants could undo their choices by clicking the “X” button next to a sentence in a list or by changing their choice in the radio buttons.

After sentences were marked, clicking the Refine Results button would display a new set of results produced through relevance feedback. The participant could then mark more sentences and repeat the process. They could also perform a search instead.

If a search was performed after a relevance feedback step, a two-step process would ensue. First, any marked sentences not already integrated into the relevance feedback model would be integrated into the query. Second, the resulting query vector would be added with equal weight to the new search query. This ensured that the relevance feedback already issued would not be lost when the user typed in a new search query.

STUDY DESIGN

Our goal was to determine whether relevance feedback was better able to support finding literary themes than keyword search alone. We decided upon a between-participants study design with a single theme from Shakespeare. Participants were randomly assigned to either the relevance feedback system or the search-only system. Both systems retrieved results from the same collection: the complete works of Shakespeare, published as electronic texts by the Internet Shakespeare Editions.

The search-only system used the same vector space retrieval model as the relevance feedback system, except that feedback was disabled – there was no "Refine Results" button, and the system did not adjust the query vector in the direction of sentences marked relevant.

Participants were shown an explanation of the theme, with two examples. Then, they were asked to find as many more examples of the same theme as they could within five minutes. When the time limit expired, the sentences that the participants had marked relevant were logged.

Finally, the participants were taken to a self-evaluation questionaire. On a scale of 1 to 5, they were asked to rate their understanding of the task, their understanding of the theme, their perception of how easy or difficult the task was, and their perception of how well they performed. Lower numbers were worse, and higher numbers were better.

We instrumented the study so that people could take it remotely (using cookies to guard against repeat participation). We logged the number of searches, number of refines, and the numbers and ID's of sentences marked relevant and not relevant.

Since we were dealing with literary themes, we decided to restrict participation to people with at least college-level backgrounds in English language or literature. Before starting the study, participants were asked to self-report their level of experience in English language or literature. Calls for participation were tweeted by the authors, and sent to mailing lists at the English departments at Berkeley and Stanford.

The theme in the study was chosen by our domain-expert collaborator, Dr. Michael Ullyot at the University of Calgary. Dr. Ullyot is a Professor in the English department, and regularly teaches Shakespeare courses. The theme he chose was "the world as a stage", in which imagery relating to actors, acting, or the theater is used in relation to life, or real-world events. The two examples he selected to illustrate the theme were:

All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances;
And one man in his time plays many parts,
His acts being seven ages.
(As You Like It, Act 2 Scene 7 Lines 139–143)

Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more: it is a tale
Told by an idiot, full of sound and fury,
Signifying nothing.
(Macbeth Act 5 Scene 5 Lines 23–27)

In the relevance feedback condition, the query vectors were automatically adjusted to include the two sentences above.

Expert Evaluation

Once all participants had finished, we submitted their sentences to our Shakespeare scholar, Dr. Ullyot. He marked each sentence either relevant or not relevant to the "world as a stage" theme. Using these scores, we were able to derive the set of 46 sentences identified as relevant across all participants and the number of relevant sentences found by each participant. We were thus able to compute precision and recall for each participant, with recall defined against the union over all the participants of sentences judged relevant by our expert.

RESULTS

23 participants with the requisite background completed the study. Of these, 11 had PhDs, 3 had Master's degrees, 6 had bachelors' degrees, and 3 were current undergraduates in English language or literature. The search-only condition received 12 participants, and the relevance feedback condition received 11.

We were interested in differences between the following observables across the two systems:

- Number of sentences found,
- Number of relevant sentences found,
- Precision,
- Recall,
- Perceived satisfaction with own performance,
- Perceived difficulty of task
- Number of searches performed

Our hypothesis was that relevance feedback would be more effective than keyword search-only at helping find examples of the theme. In terms of our observables, this implied the following changes: more sentences, and more relevant sentences, higher precision and recall, higher task satisfaction (with no increase in perceived difficulty), and fewer searches.

To test our hypothesis, we performed a two-sided Wilcoxson rank sum test on each of the above observables across the two conditions. We chose this test because we did not have paired samples, and we could not assume that the observations were normally distributed.

Our results were suggestive, but not statistically significant. For all observables, the average values in the relevance-feedback condition differed from the average values in the search condition in the direction consistent with our hypothesis. On average, participants in the relevance feedback condition found more sentences, of which more were relevant, and performed fewer searches in order to do so (Figure 2). This resulted in higher precision and recall (Figure 3). They also had higher task satisfaction, with no change in perceived difficulty (Figure 4).

However, none of the differences were statistically significant, and the confidence intervals on the differences were so broad as to render the study inconclusive. Larger sample sizes would have reduced this uncertainty.

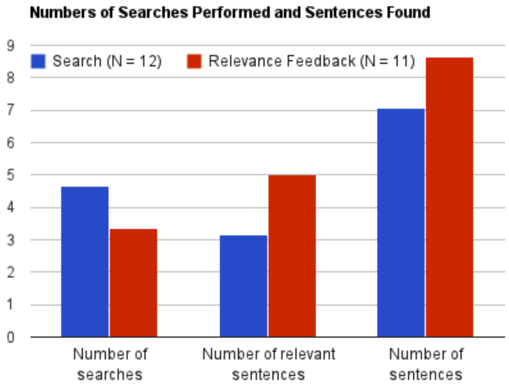


Figure 2. A comparison of the average number of searches performed, number of sentences found, and the number of relevant sentences found across the two conditions, relevance feedback (red) and keyword-search only (blue). The differences are consistent with our hypothesis that relevance feedback (red) makes it easier to find relevant sentences. With relevance feedback, there are fewer searches, but more sentences, and more relevant sentences.

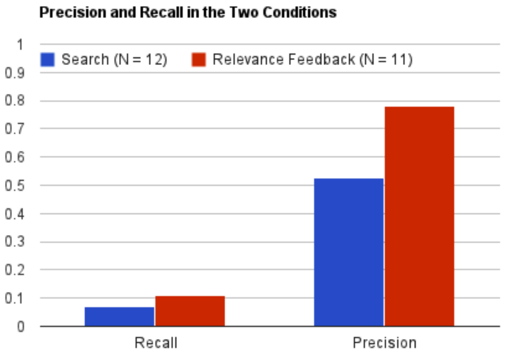


Figure 3. A comparison of average precision and recall across the two conditions, relevance feedback (red) and keyword-search only (blue). The increased precision and recall in the relevance feedback condition (red) is consistent with our hypothesis that relevance feedback is an effective aid for finding literary themes.

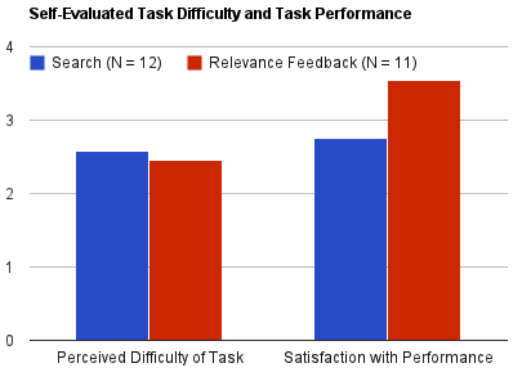


Figure 4. A comparison of the average self-evaluated task difficulty and performance satisfaction across the two conditions, relevance feedback (red) and keyword-search only (blue). Participants in the relevance feedback condition (red) were more satisfied with their performance, consistent with our hypothesis that relevance feedback makes it easier to find relevant sentences.

DISCUSSION AND FUTURE WORK

Our results suggest that relevance feedback was more effective than keyword search alone for finding examples literary themes. We intend to conduct a larger study to establish statistical significance. In this second study, statistical power could be increased by collecting paired samples in which each participant does two different theme finding tasks, one on the relevance feedback and one on the search-only system.

There are also improvements that can be made to the relevance feedback system. The first is the size of the retrieved units. When we described the system to our literary-scholar collaborators, a frequent objection was that sentences were an unnatural unit when looking for themes. However, paragraphs might be too big or too small, depending on context. To address this problem, we could segment the text into consecutive topically-coherent units using an approach such as TextTiling [3]. Instead of retrieving sentences or paragraphs, we could instead retrieve these units.

Another area for improvement is the feature-space representations of the units. With syntactic parsing, we could extract subject-object and dependent-modifier relationships between words, and incorporate them into the feature vectors. Synonymy could also be employed as a way to broaden the query. The presence of a word in a sentence feature could “activate” (with some diminished weight) the features for all the synonyms of that word.

ACKNOWLEDGMENTS

This work is supported by grants HD-51244-11 and HK-50011 from the National Endowment for the Humanities.

REFERENCES

1. Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proc. ACM Conference on Information and Knowledge Management*, Association for Computing Machinery (Lisbon, Portugal, 2007), 213–222.
2. Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. The vocabulary problem in human-system communication. *Communications of the ACM 30*, 11 (1987), 964–971.
3. Hearst, M. A. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23*, 1 (Mar. 1997), 3364.
4. Hearst, M. A. Chapter 3 models of the information seeking process. In *Search User Interfaces*, 1st ed. Cambridge University Press, New York, NY, USA, Sept. 2009, 64–90.
5. Heilman, R. B. *Magic in the Web: Action and Language in Othello*, first edition ed. University of Kentucky, 1956.
6. Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. International Conference on Intelligence Analysis*, vol. 1 (MacLean, VA, USA, 2005), 2–4.
7. Rocchio, J. J. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, G. Salton, Ed., Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, NJ, USA, 1971, ch. 14, 313–323.
8. Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*, 5 (1988), 513–523.
9. Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. In *Readings in information retrieval*. Morgan Kaufmann, 1997.