A Retrieval System based on Sentiment Analysis

Wei Zheng
University of Delaware
zwei@udel.edu

Hui Fang
University of Delaware
hfang@udel.edu

ABSTRACT

The aim of HCIR Challenge is to encourage systems that can help users to quickly find the needed information and understand the meanings of the retrieval results. We participated in the second task of HCIR Challenge that helps users understand the competing perspectives on controversial queries. The reason for us to choose this task is that users need to know the different opinions on the controversial queries while it is difficult for them to get the information in traditional retrieval systems that just return a list of documents without further analyzing the knowledge in these documents. We develop a system that returns documents based on their sentiments and topics given the query. The system retrieves documents given the original query and displays the topics in the results that are for or against the query. It then connects the pair of similar topics belonging to the for and against categories. Therefore, users can easily know the perspectives in the query and compare the positive and negative arguments discussing these perspectives in our system instead of reading all the returned documents and summarize those information by themselves in traditional retrieval systems.

1. INTRODUCTION

The HCIR Challenge encourages the novel systems that give users more guidance to quickly find the needed information and get more knowledge from the retrieval results. It has three tasks that include learning about the topic having a long history, understanding the controversial perspectives on the controversial query and answering the question requiring looking at more than one document. We participated in the second task because users need to understand the retrieval results and know different arguments in the results while they cannot easily get these information from the traditional retrieval systems that simply return a list of documents as the results without giving any further knowledge of the results. Therefore, it is necessary to re-organize the retrieval results and provide facilities for users to understand the results.

We developed a system that returns documents given the query according to their sentiments and topics. It not only returns relevant documents but also displays the perspectives in the query and different arguments in each perspective. The system was based on the Lemur toolkit which we added a re-organizing component to. The system retrieves the documents given the query with Lemur and classifies the returned documents to the categories that are for or against the query according to their sentiments analysis results using the toolkit OpinionFinder. It then uses the probabilistic latent semantic analysis (PLSA) [1] algorithm to extract the topics in the documents of each category and displays the topics with both the topic descriptions and the documents belonging to the topics. We also connect the similar pair of topics belonging to different categories. Each pair of similar topics corresponds to a perspective of the query and each topic correspond to the arguments for or against that query perspective. The result of the system allows user to easily know the perspectives in the query and compares the opinions in each perspective by reading the arguments of the perspective.

The paper is organized as follows. We describe the framework of the system in Section 2 and show the system interface and results in Section 3. We then conclude in Section 4.

2. THE SYSTEM FRAMEWORK

As shown in Figure 1, the steps of the system are listed as follows:

1. Retrieving documents. We use the Lemur toolkit to build the index of New York Time corpus and retrieve documents with Dirichlet prior retrieval function.

2. Analyzing the sentiment of the returned documents. We use the OpinionFinder toolkit to analyze the sentiments in the documents returned in the above step and classify the documents into categories that are for or against the query.

3. Mining the topics of documents in each category. We extract the topics of documents in each category with PLSA topic modeling method using Lemur toolkit. Each document and term is assigned to the topic in which they have the highest probabilities. The number of topics in each category is set to be 5.

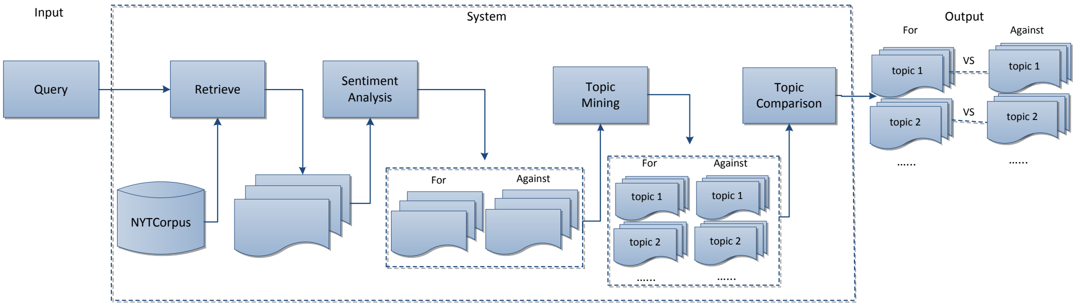


Figure 1: the architecture of our system

4. Comparing topics belonging to different category. After modeling the topics in the for and against category, we use cosine similarity method to compute the similarity between topics using their terms. The pair of topics belonging to different categories is judged to be similar and correspond to one perspective of the query if their similarity is larger than a threshold which is set to be 0.2.

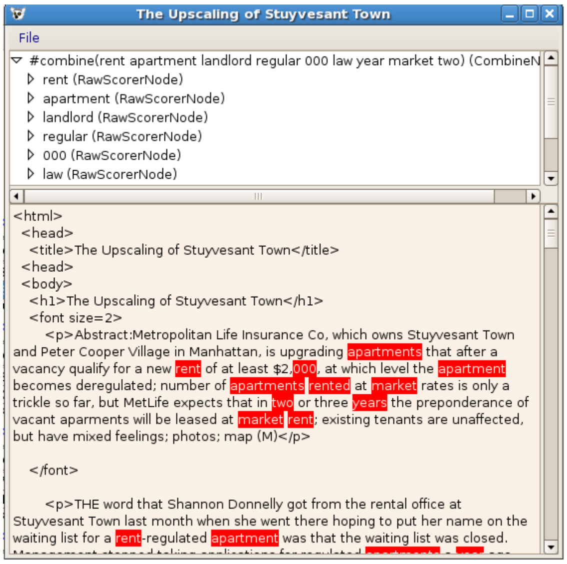
3. THE SYSTEM INTERFACE

Figure 2 shows the retrieval results of the query *rent control in New York*. The system uses PLSA algorithm to extract the topics in the documents of each category. The default number of topics is set to be 5. The interface displays the top ranked terms of the topics and the information of related documents in each topic. We also build connections between the similar topics belonging to different categories and use *VS* to show that two topics are discussing the same query perspective but expressing different opinions. As shown in Figure 2, topic 0-3 in the *FOR* and *AGAINST* categories are discussing the same perspectives while topic 4 in the two categories are discussing different perspectives. We can see some encouraging results in the topics. For example, the topic 0 of *FOR* and topic 0 of *AGAINST* are different opinions about the influence of the rent control to the rent market, and topic 1 on both sides are about the change of the rent cost.

Users can check the detail results of each topic when clicking the *detail* button. The table in each topic shows the ID and title information of the documents. Users can see the original document with highlighted topic terms, as shown in Figure 3, when clicking the row of the document.

4. CONCLUSION

The traditional retrieval systems just return a list of documents for users to read. It is difficult for users to know all the perspectives in the query. We developed a system that returns documents based on their sentiments and topics. Therefore, users can easily know the topics that are for or against the query when reading the topic description and related documents. They can also compare different arguments on the same perspective when comparing the similar topics displayed in the system. There are some interesting future works for the system. First, we can analyze the documents in each sentiment category to automatically decide the number of subtopics. Second, we can extract the terms that express the common meaning of the pair of similar topics as the description of that pair and use terms expressing opposite opinions as the description of the individual topic in the topic pair.



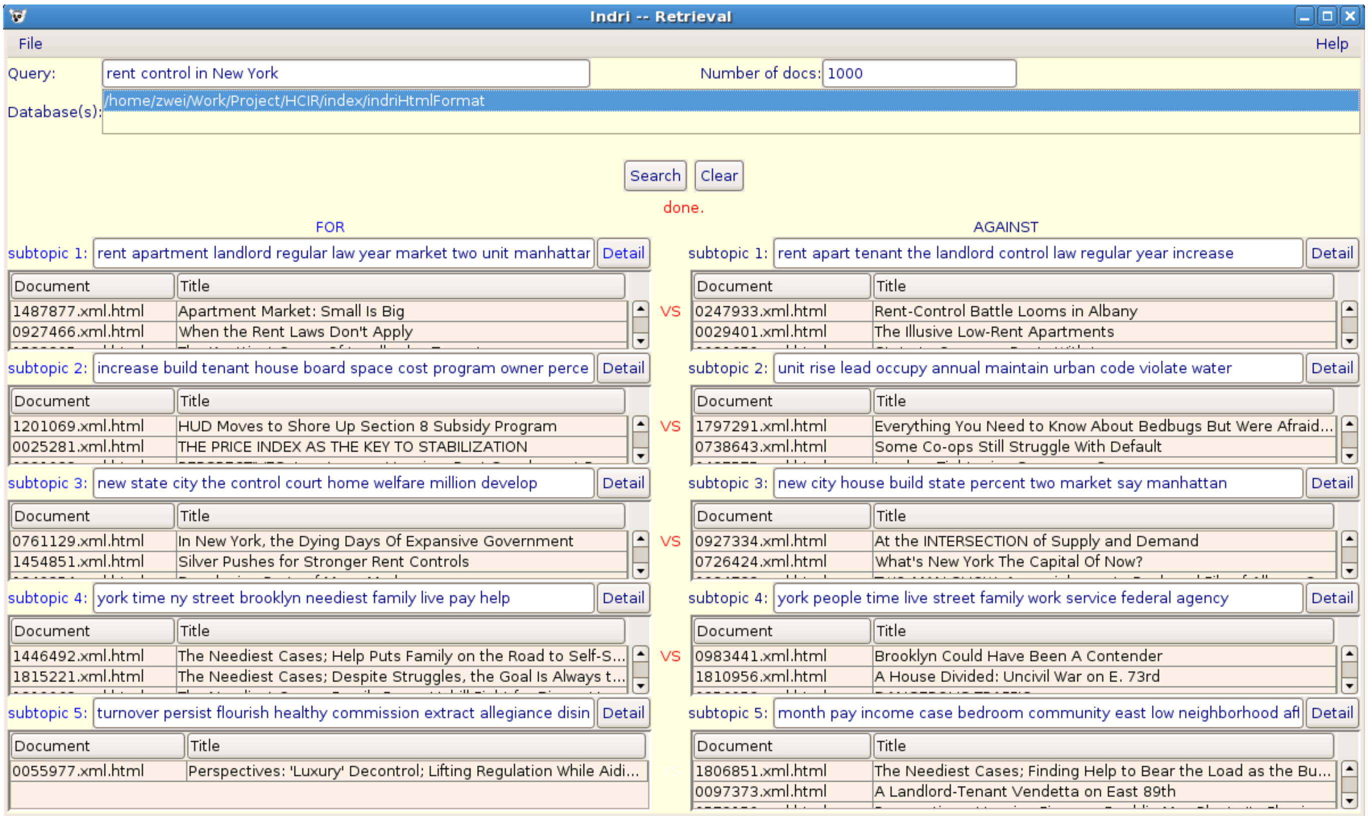
```
▽ #combine(rent apartment landlord regular 000 law year market two) (CombineN
  ▷ rent (RawScorerNode)
  ▷ apartment (RawScorerNode)
  ▷ landlord (RawScorerNode)
  ▷ regular (RawScorerNode)
  ▷ 000 (RawScorerNode)
  ▷ law (RawScorerNode)
```

```
<html>
 <head>
  <title>The Upscaling of Stuyvesant Town</title>
 <head>
 <body>
  <h1>The Upscaling of Stuyvesant Town</h1>
  <font size=2>
    <p>Abstract:Metropolitan Life Insurance Co, which owns Stuyvesant Town
and Peter Cooper Village in Manhattan, is upgrading apartments that after a
vacancy qualify for a new rent of at least $2,000, at which level the apartment
becomes deregulated; number of apartments rented at market rates is only a
trickle so far, but MetLife expects that in two or three years the preponderance of
vacant aparments will be leased at market rent; existing tenants are unaffected,
but have mixed feelings; photos; map (M)</p>

  </font>

    <p>THE word that Shannon Donnelly got from the rental office at
Stuyvesant Town last month when she went there hoping to put her name on the
waiting list for a rent-regulated apartment was that the waiting list was closed.
```

Figure 3: The display of the original document

5. REFERENCES

[1] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of UAI'99*, 1999.



Query: rent control in New York

Number of docs: 1000

Database(s): /home/zwei/Work/Project/HCIR/index/indriHtmlFormat

done.

FOR

subtopic 1: rent apartment landlord regular law year market two unit manhattar

Document	Title
1487877.xml.html	Apartment Market: Small Is Big
0927466.xml.html	When the Rent Laws Don't Apply

subtopic 2: increase build tenant house board space cost program owner perce

Document	Title
1201069.xml.html	HUD Moves to Shore Up Section 8 Subsidy Program
0025281.xml.html	THE PRICE INDEX AS THE KEY TO STABILIZATION

subtopic 3: new state city the control court home welfare million develop

Document	Title
0761129.xml.html	In New York, the Dying Days Of Expansive Government
1454851.xml.html	Silver Pushes for Stronger Rent Controls

subtopic 4: york time ny street brooklyn neediest family live pay help

Document	Title
1446492.xml.html	The Neediest Cases; Help Puts Family on the Road to Self-S...
1815221.xml.html	The Neediest Cases; Despite Struggles, the Goal Is Always t...

subtopic 5: turnover persist flourish healthy commission extract allegiance disin

Document	Title
0055977.xml.html	Perspectives: 'Luxury' Decontrol; Lifting Regulation While Aidi...

AGAINST

subtopic 1: rent apart tenant the landlord control law regular year increase

Document	Title
0247933.xml.html	Rent-Control Battle Looms in Albany
0029401.xml.html	The Illusive Low-Rent Apartments

subtopic 2: unit rise lead occupy annual maintain urban code violate water

Document	Title
1797291.xml.html	Everything You Need to Know About Bedbugs But Were Afraid...
0738643.xml.html	Some Co-ops Still Struggle With Default

subtopic 3: new city house build state percent two market say manhattan

Document	Title
0927334.xml.html	At the INTERSECTION of Supply and Demand
0726424.xml.html	What's New York The Capital Of Now?

subtopic 4: york people time live street family work service federal agency

Document	Title
0983441.xml.html	Brooklyn Could Have Been A Contender
1810956.xml.html	A House Divided: Uncivil War on E. 73rd

subtopic 5: month pay income case bedroom community east low neighborhood afl

Document	Title
1806851.xml.html	The Neediest Cases; Finding Help to Bear the Load as the Bu...
0097373.xml.html	A Landlord-Tenant Vendetta on East 89th

Figure 2: The retrieval results