CS547-Homework1

Peng Li

January 28, 2016

1 For example: we search some news about football match, this is not ad-hoc retrieval. The problem is that may be new documents just come in when we already got our retrieval result and we dont know is the new document is better.

2 For example, there is a news from web page

```
<article id="article" class="article">
<header><div class="byline">
<span class="source">AP</span>
<span class="time">January 24, 2016, 10:04 AM</span></div>
<h1 class="title" itemprop="headline">Flight 370 speculation
surrounds Thai beach debris</h1>
</header><div class="article-image" itemprop="image">
<p><strong>BANGKOK </strong>— A large chunk
of metal that could be from an aircraft washed ashore in southern
Thailand, but Malaysian authorities on Sunday cautioned against
speculation of a link to a <a href="http://www.cbsnews.com/
malaysia-airlines-flight-370/" target="_blank">Malaysia Airlines
Flight 370</a> missing almost two years.</p><p>Flight 370 lost
communications and made a sharp turn away from its Beijing
destination before disappearing in March 2014. It is presumed
to have crashed in the Indian Ocean, and only one piece of debris
has been identified as coming from the plane, a slab of wing that
 washed ashore on Reunion Island in the western Indian Ocean last
 July.</p><p>Malaysian Transport Minister Liow Tiong Lai said he
 instructed Malaysian civil aviation officials to contact Thailand
 about the newly found wreckage, a curved piece of metal measuring
 6.5 feet by 10 feet with electrical wires hanging from it and
 numbers stamped on it in several places.</p><p>"I␣urge␣the␣media
␣and␣the␣public␣not␣to␣speculate␣because␣it␣will␣give␣undue␣pressure
␣to␣the␣loved␣ones␣of␣the␣victims␣of␣flight␣370," he said.</p>
 <p>Thailand's Transportation Ministry said four Malaysian officials
 and two Thai experts will visit the site Monday.</p></div>
</article>
```

I focus on MH370 and my query is ”MH370 aeroplane crash”. Only number ”370” appears in this article. So, if we use all word in this html file when we preprocess, we will not get this article or this article will have a low rank among all result although this article has a high relevance with my searching key word. If this article is preprocessed based on structural properties, I will get better result.

3.1 controlled vocabulary - recall. Indexing text with controlled vocabulary is easy to cross vocabulary gap which means user will get more relevant documents and it will improve recall, but maybe get more irrelevent documents. For example: my query is vehicle and I can get many documents with key word car and auto, but also can get excipient.

3.2 remove stopwords - both When I search Purdue,I will not get a lot of result about the or is although this word appear more times than other words because stopwords are removed. So I think stopwords removing will improve both precision and recall.

3.3 stemming - recall Stemming can match word better but it also have disadvantage. When I search American and I want to know more about America people, may be I will get lots of results about America. The precision will be decreased but I think I will get more relevant document and it can improve recall.

3.4 Representing phrases as a singer term - precision When I query hit a home run, I will get many results related to baseball because hit a home run in there is a baseball term rather than other things, such as ”home”,”hit”,”run”.

4 In most case, we get many results means these results have a low precision and a high recall. This is because the more results we get, the more relevant results we get. So we can get a high recall. But we get more wrong result at the same time. So we get a low precision. If we get less results and they are relevant, we can get a high precision. Although we have a high precision, we have a low recall because we get only little part of all relevant results. So, there is a trade off between precision and recall. In that case, Balanced F-measure will not have much change for different information retrieval system. It can’t correctly evaluation an information retrieval system. Thats why balanced F-measure is not good for information retrieval system evaluation.

5.1

Term ID	Term	Documents
1	Purdue	1,3,4
2	Information	1,4
3	Retrieval	1,2,3,4

5.2

"Purdue Information" : $\vec{q_1} = (1{,}1{,}0)$
Doc3 : $\vec{q_2} = (2{,}0{,}3)$
$cos(\vec{q_1}, \vec{q_2}) = \frac{1\times2+1\times0+0\times3}{\sqrt{1^2+1^2+0^2}\sqrt{2^2+0^2+3^2}} = \frac{\sqrt{26}}{13}$
when I am search "Purdue Information", I think Doc1 will be ranked higher

5.3

I think Doc4 is most likely about our class "Information Retrieval"