Report:

Information Search Engine: Diversity

ZHANG Xiao

Abstract

The traditional information research engine consider the documents are independent each other, the relevance of the information from one document depends on what is already know about the subject and in turn affects the relevance of other documents subsequence. This mini project concrete on the analyse of diverse information search methods, I firstly implement a benchmark without consider the sub-subject, then some diversity algorithms like clustering, greedy ranking are introduced. In addition, the evaluation methods are P@n, CR@n.

key words: information search ; diversity ; clustering, MMR

Introduction

The traditional information search engine take relevance as an important factor of evaluation, but in reality, diversity is an element we can't ignore. The diversity in information research system can be consider as three different definitions: the diversity in the request addressed; the diversity in the reference discipline used to account for research result; the diversity in measure methods used to sort the result. Because the existence and evidence of diversity in information research engine, the traditional search algorithm can not promise a relative high relevance result.

The diversity of information search engine can be applied to lots of scenarios, recommendation system for example, Most of the diversity method applied on the sorted result is somehow not always useful. Calculate Complexity and time complexity, memory required. Even thought, we know the sub topics, it is always difficult to find a optimal ranking of result.

So the goal of this project is to illustrate the importance of diversity in a search engine and to implement some classic method to improve the diversity of basic search result. In this article, I begin by the introducing the exist problem in traditional search engine. Next, some replace algorithm considering the diversity like clustering will be involved. I then evaluate and compare the performance of the benchmark and the proposed algorithm on the easy dataset, using different measure metrics. At last, I propose some extensions to my current work in the last chapter.

State of Art

Two main categories are proposed to solve the diversity problem, which are : Clustering and Greedy algorithm[1] .

- Clustering to improve the diversity

  Post-retrieval clustering

  Three step:

  1. Obtain a ordered document searched by traditional search engine
  2. Cluster with top N documents
  3. Resort the document by considering the diversity of results.

- Greedy Algorithm
  As the difference method to represent the similarity between two query or documents, the greedy algorithm can detailed as MMR(Maximal Marginal Relevance)[2], DM, Maximum the minimum distance.
  Greedy can easily reach a local optimal performance, but not promise a global opt Imation. A Dynamic Programme method is proposed two realize at the same time the optima ion and programing to gain a better performance.

Proposed approach

- Clustering methods:
  Two kinds of Clustering methods are introduced in this project, One is a partitioning method: kmeans, the other is Hierarchical method: Birch.
  K-Means :
  K-Means Clustering is one of many techniques within unsupervised learning that can be used for text analysis, it is known to be efficient in clustering large data sets. The K-Means algorithm aims to partition a set of objects, based on their attributes/features, into *k* clusters, where *k* is a predefined or user-defined constant. The main idea is to define *k* centroids, one for each cluster. In the text clustering, we firstly represent documents with TF or TFIDF, then apply the K-Mean cluster method.
  K-means is an efficient method but the result of Kmeans is very sensitive to the initialization.

  Birch:
  The Birch builds a tree called the Characteristic Feature Tree (CFT) for the given data. The data is essentially lossy compressed to a set of Characteristic Feature nodes (CF Nodes). The CF Nodes have a number of sub clusters called Characteristic Feature sub clusters (CF Sub clusters) and these CF Sub clusters located in the non-terminal CF Nodes can have CF Nodes as children. It works well with low volume data. Hierarchical Clustering can give different partitionings depending on the level-of-resolution we are looking at. Birch clustering can be slow because it has to make several merge/split decisions.

  However, no clear consensus on which of the two produces better clustering.

- Greedy ranking Algorithm
  A general version of greedy algorithm is :

```
Inputs: Set of unranked documents U;
        ranking size K
        for i = 1, 2, . . . , K do
                $di = \arg\max_{d \in U} \textit{value} (d; d1, . . . , di-1)$
                $U = U - \{di\}$
                endfor
        return the ranking $<d_1, . . . , d_K>$
```

The key elements here is to appropriately define the “value” function, in this project, I use the maximal Marginal relevance proposed in the paper of Carbonell and Goldstein to define the value. A document has a high marginal relevance if it both relevant to the query and contains minimal similarity to previous documents.

MMR:

$$value_{\mathrm{MMR}}(d; d1, \ldots, di-1) = \alpha\ \mathrm{Sim}_1(d, Q) - (1 - \alpha) \max \mathrm{Sim}_2(d, d_j)$$

Experiment

- Benchmark
- Measure Metrics (precision recall, mean precision, P@n, CR@n)

  P@n: Precision of n documents

  It is a method to evaluate the relevance of a information search system, which aimes to calculate the rate of relevant document in the n observed documents.

  $$\mathrm{P@n}=\frac{number\ of\ relevant\ document}{n}$$

  CR@n: Cluster Recall of n documents is a classic method to evaluate the diversity. It’s purpose is to display the rate of diffrents sub topic in the observed results.

  $$\mathrm{CR@n}=\frac{number\ of\ sub\ topic\ in\ the\ first\ n\ documents}{Total\ number\ of\ sub\ topic}$$

  Generally, we take 10 or 20 for the number of documents, because it corresponds approximately to the number of images that can be shown in one web page. So, in my experiment, I use P@20 and CR@20 as the evaluate method.
- Pre-processing of data

  The dataset: easyCLEF08

  Basic information search model: Vectorial model

  Even thought many information search model are already implemented such as Okapi, PageRank, Hits. But, in the experiment, I use Vectorial model as the basic model, Because it can promise a relative good performance with less exultation times.

  Index and representation method: Different text representation methods are used in the experiment, such as TF(Term Frequency), idf(Term Frequency-Inverse).
- Results

  Clustering:

  K-mean Clustering:

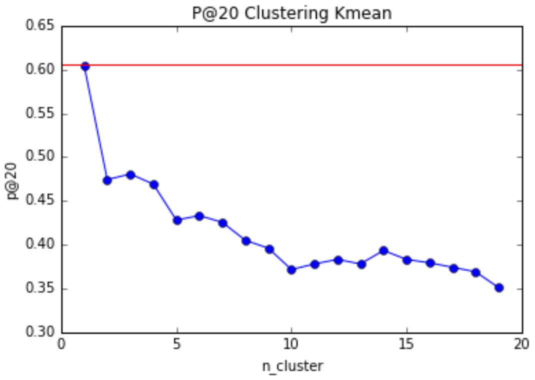

Fig 1 P@20 k-means clustering

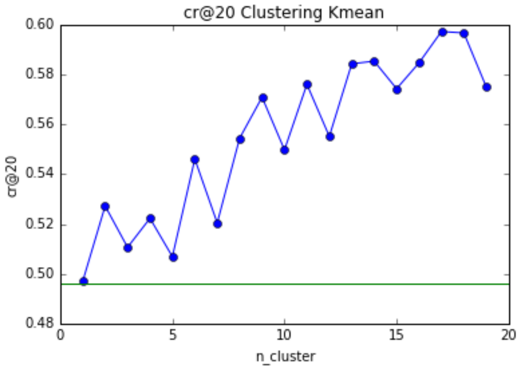

Fig 2 CR@20 k-means clustering

Birch Clustering:

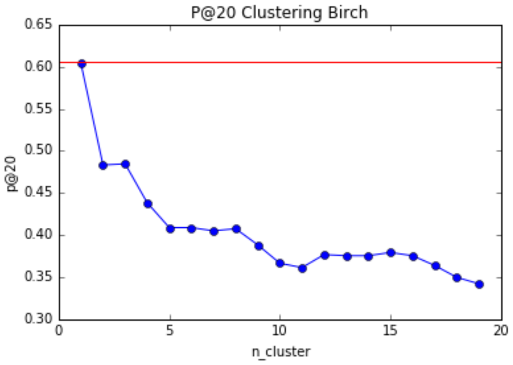

Fig 3 P@20 Birch clustering

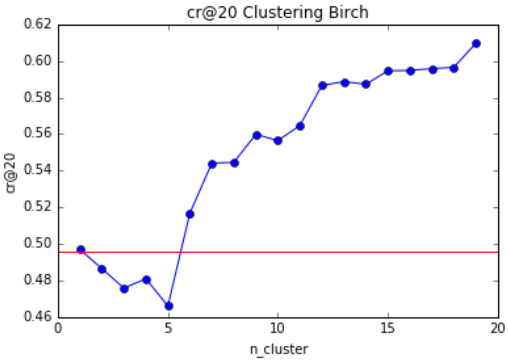

Fig 4 CR@20 Birch clustering

Greedy method: MMR:

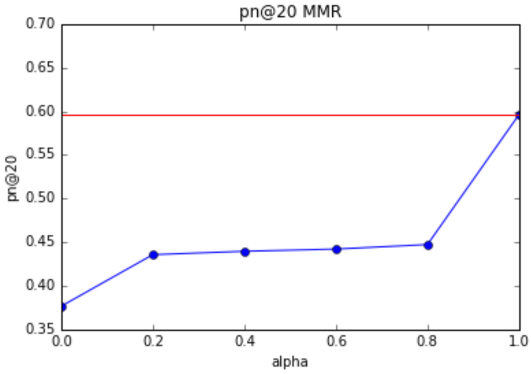

Fig 5 P@20 MMR Greedy

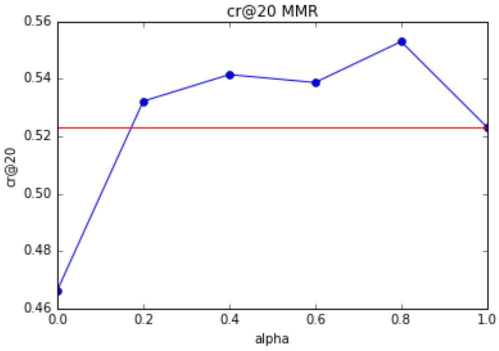

Fig 4 CR@20 MMR Greedy

From the figure, We can observe that, with the Clustering method, when CR@20 increase, the p@20 will decrease which means that clustering method scarify the relevance to improve the diversity. With the rise of number of cluster, the CR@n value improve with a crease.

As to the MMR method, when the alpha equal to zero, both cr@20 and p@20 reach the minimum, with the increase of value alpha, pn@20 increase too, reach the niveau of base line in the final, but cr@20 increase at the very first, and decrease until the baseline finally.

- Comparison

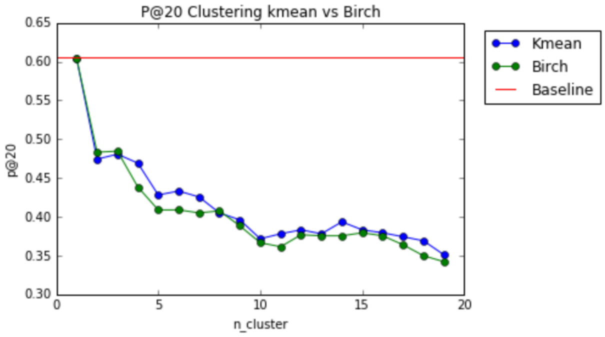

Fig 7 P@20 Birch vs K-means

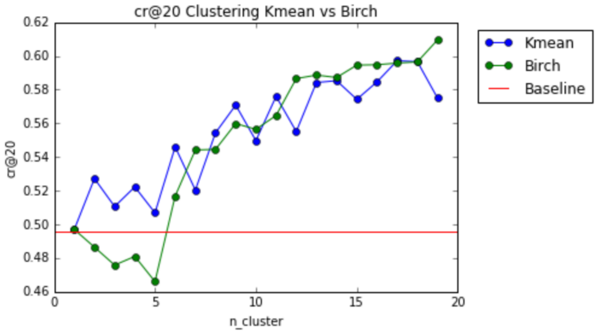

Fig 8 CR@20 Birch vs K-means

If we compare the two different clustering method: K-Means and Birch, We can observed that cr@20 diversity of Birch is slightly better than means, but relevance is less good than k mean.

Conclusion

In this project, I introduced two types methods to improve the diversity of information search results: Clustering and Greedy method. For the Clustering method, I used K-Means and Birch methods as two different approaches for documents clustering, and compared the performance of those two. The actual method scarify the relevance to improve the diversity, so in the future, we can try to develop a innovate method to improve the diversity and keep the relevance at the same time.

Bibliography

[1] The use of hierarchical clustering in information retrieval, Jardine and van Rijsbergen, information storage and retrieval, 1971.

[2] Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, Zhai et al., ACM SIGIR, page 10-17, 2003

[3] Hierarchical clustering diversity pseudo-relevance feedback for social image search result diversification. Boteanu et al., CBMI, page 1-6, 2015.

[4] The use of MMR, diversity based reranking for reordering documents and producing summaries. J. Carbonell and J. Goldstein, Proceedings of SIGIR 1998, 1998.

[5] Benchmark, easyCLEF08

[6] K-means, Birch, Sklearn http://scikit-learn.org/stable/modules/clustering.html