IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-4 E-ISSN: 2347-2693

Pattern Based Frequent Term Retrieval Search Using Text Clustering

R.Krithika[1], G.Sathish Kumar[2]

[1]*M.Tech Scholar, Department of Computer Science & Engineering, MNSK College of Engineering, Pudukkottai*
[2] *Head, Department of Computer Science & Engineering, MNSK College of Engineering, Pudukkottai*

www.ijcseonline.org

Received: Mar/23/2016 Revised: Apr /03/2016 Accepted: Apr/19/2016 Published: Apr/30/2016

Abstract— Clients are known to experience troubles in dealing with information retrieval look outputs, particularly if those yields are above a certain size. It has been contended by several analysts that look yield Clustering can help clients in their collaboration with IR frameworks in some retrieval situations, providing them with an review of their results by abusing the topicality information that resides in the yield but has not been used at the retrieval stage. This review might enable them to find applicable records more effortlessly by focused on the most promising clusters, or to use the Groups as a starting-point for question refinement or expansion. In this paper, the results of tests carried out to assess the viability of Clustering as a look yield presentation technique are reported and discussed.

Keywords— *Content Clustering, Pattern Mining, Content Retrieval, Clustering Algorithm.*

I. INTRODUCTION

A user's collaboration with a look yield is frequently far from optimal. Particularly when the yield exceeds a certain threshold, clients are inclined to sample just a few records or abandon the question altogether. With a Boolean system, the look yield can be reduced by introducing extra look terms, but studies appear that a great majority of clients do exceptionally little or no Boolean searches. Indeed experienced clients may not be willing or able to find the fitting terms to narrow down a search.

Most IR frameworks give the clients with a relevance-positioned list to help them find applicable records easily, but in cases where a client experiences troubles in expressing his information need, or has a more exploratory approach towards a look output, or when numerous records have the same score, significance positioning may not be exceptionally helpful. It has been proposed that Clustering can help clients in such cases, by showing them some kind of pattern existing in the record set, enabling them to review the set quickly and make judgements on Clusteres of records simultaneously. Alternatively, if the coverage of the yield is infitting for the user's need, the topicality information presented in the Cluster representations may give cues for modifying the query. If Clustering achieves a accommodating categorisation of the documents, it may moreover act as a versatile question extension aid.

This article reports test discoveries from a PhD project, in which yield Clustering was investigated as a client collaboration tool, working on a constrained number of records recovered by the Okapi probabilistic look engine. Can&Ozkarahan's C³M calculation was used with little alterations for clustering the documents, and probabilistic techniques were utilized for record retrieval and Cluster representations.

II. CHOICE OF METHOD

There are numerous diverse Clustering techniques which are neither mutually exclusive nor can be neatly sorted into a few groups. Progressive techniques are the broadest family to be sorted under one group; see Everitt for a detailed classification. In the past, much dialog on the choice of Clustering techniques has centered on processing efficiency, as Clustering algorithms are notoriously complicated and processor-intensive. Nowadays this is of much less concern, and the different techniques can be judged purely on their results.

No Clustering technique can be judged 'best' in all circumstances, and it is infeasible to comprehensively test a wide range of Clustering techniques before picking one. In this project, the C3M technique was picked as it conformed to the theoretical soundness criteria, and given a implies of estimating the optimum number of Groups as well as recognizing Cluster centroids and forming Groups around the centroids. It moreover had a history of great execution with the Inspec database which was available for the venture experiments.

Unlike progressive Clustering techniques which have been generally favoured by CBR researchers, this technique allowed overlapping; a feature that was found to be accommodating in the connection of this project. In contrast

to the general inclination for progressive Clustering techniques in CBR, it is contended here that record collections do not have any intrinsic qualities that make them fitting for hierarchic conceptualisation. On the contrary, due to the assortment of viewpoints a record may cover, an covering grouping where a record can be a member of more than one Cluster appears more appropriate. Everitt's statement beneath might apply to record Clustering as well:

"It is in biological applications such as the evolutionary trees that progressive classifications are most relevant. Progressive Clustering procedures are, however, now used in numerous other fields in which progressive structures may not be the most appropriate. The danger of imposing a progressive plan on information which is essentially non-progressive is clear."

It is puzzling to observe CBR researchers' overwhelming inclination for progressive methods, particularly when we consider that they are more requesting in their assumptions than their non-progressive counterparts. Progressive techniques attempt to create accommodating Clusteres out of a set of items at different diverse levels: for a record set of 100 documents, this implies that the records can be divided into, say, 2, 4, 7, 10, 15 and 30 Clusteres in a huge way.

In fact, relook work to date shows that indeed when a progressive structure is preferred, it is typically only the bottom-level Groups which give great retrieval results. Indeed if we have a progressive representation, we do not seem to have much use for its upper levels. This is entirely intuitive when we consider the amount of information lessening included there.

For this project, the maximum size of the record sets to be Clustered was set at 50. This figure was extensive enough to make it worthwhile to apply clustering, and little enough to Cluster without excessive information reduction. In these circumstances we required nothing more than a simple partitioning method, making the use of the progressive technique indeed less appropriate.

III. EXISTING METHOD INCLUDED IN YIELD CLUSTERING

When implementing a Clustering method, it is first essential to choose on the sort and number of variables to be used. Records are generally represented by terms for Clustering purposes, but it has been contended that significance can't be constrained to topicality; and a assortment of other elements such as authors, journal, obtainability, cost, and previously seen records all affect client decisions. In this implementation, the record representation was constrained to term occurrences as most of these other elements were entirely troublesome to measure and incorporate in the implementation.

As Clustering results would be affected by the number of terms used, different tests were done to find the estimated number of terms that could be anticipated to produce a adjusted dispersion of records among clusters, and set upper and lower boundaries for the algorithm. The lower limit was set at four, since terms occurring only in one or two records would not be of any use at all.

Setting the upper limit included more thought. Obviously the more terms are used the more information is available for the Clustering algorithm, but the greater the hazard of obscuring the Cluster structure. One technique to reduce this hazard was to use a list of about 900 stopwords in conjunction with a stemming algorithm, to weed out non-contextual terms. In addition, a long list of synonyms (about 950) was utilized in request not to double-count words that could be used interchangably, or to let similarities between records reprimary unexploited due to diverse wordings of the same expression or idea.

The complete number of terms was then constrained to a maximum of 70, but after applying the above techniques the number of eligible candidates usually fell somewhat short of this figure. In cases where it was exceeded, the candidate terms with highest Term Choice Esteem were chosen. TSV was a more fitting standard for our purposes than term weight, since it correlates with a term's capacity to discriminate between record groups.

It was moreover essential to choose whether to use the calculation in weighted or binary mode. Binary mode was picked as it was simpler to implement, and C3M was reported to perform well with it.

Another Cluster of choices included producing concise representations of Cluster topics for users' viewing. Such representations could in principle consist of delegate terms and / or delegate titles, and it was essential to establish how to select those which would give the best implies of discrimination, and enable clients to assess the Groups most easily.

After some experiments, it was chosen to present a combination of three titles and up to ten terms, i.e. those with the highest TSVs. (For an illustration see Table 1 below.) In request to maximise discrimination, only those delegate terms that did not occur in any other Cluster representation were selected for display. Delegate terms on their own were found to be somewhat cryptic for members to use in evaluating clusters, so, in request to appear them within a huge context, it was chosen to pick those delegate

titles containing the most delegate terms, as well as listing the terms themselves.

The elective to this choice standard was to pick delegate titles based on their comparability to the Cluster seed document, ensuring that the titles of the most typical individuals of the Groups would be shown, regardless of whether they conveyed a great representation of the record contents.

In request to test these two elective approaches, ten preliminary tests were performed. These uncovered that clients found the titles to be more essential than the terms in surveying clusters, and that picking the titles according to the number of question and delegate terms they included produced a viable representation. However there was no huge distinction between the two title choice techniques in terms of client preference, and a choice was made to use both in subsequent experiments. Clients would be inquired to rank the Groups in each representation separately, and, at the end of the experiment, to compare them in their perceived usefulness.

Table 1 : Illustration Cluster representation (Query: IR, system, evaluation, performance, compare, criteria, comparative)

….
CLUSTER 2 (incorporates 15 documents) RANK( )
DELEGATE DOCUMENTS
5: A critical examination of review and exactness as measures of retrieval framework performance
17: On probabilistic notions of exactness as a capacity of recall
39: Usage and assessment of a significance criticism device based on neural networks.
DELEGATE TERMS
probabilistic - record - framework - question - examination - applicable - retrieval - assessment computerized information retrieval – accuracy
CLUSTER 3 (incorporates 9 documents)
DELEGATE DOCUMENTS
47: Comparable modeling and assessment of CC-NUMA and COMA on progressive ring architectures.
18: Latency examination of CC-NUMA and CC-COMA rings
11:Newton: Performace change through comparable analysis

IV. TEST SETUP

The primary purpose of the tests was to discover whether Clustering could be superior to significance positioning as a look yield presentation method. As significance positioning could not co-exist with a Clustering scheme, it was essential to assess the execution of Clustering against positioned retrieval before being able to propose it as an alternative. The speculations were:

Null speculation : The exactness of significance positioning can't be improved by used Clustering as a look yield presentation method.

Elective speculation : Clustering look yield can improve the exactness of positioned retrieval by creating recognizable "applicable clusters" that include fundamentally higher proportions of applicable records than the positioned yield at tantamount edge levels.

A complete of 85 client experiments, based on users' own information needs, have been led to test these hypotheses. After the first 20 experiments, execution results and client criticism were evaluated to find out ways to improve the implementation. As a result of this evaluation, some alterations were made in Cluster and record representations and ten tests were led to compare two elective techniques for selecting delegate titles as described above. Finally, 55 tests were led to attain factually huge results. The results from those tests are reported in this paper.

The first set of tests were factually inconclusive and as the usage and the test set-up were somewhat modified afterwards, their results have not been consolidated with the final results.

The general flow of the tests was as follows:

Clients were inquired to write down their information need (question terms) in a pre-questionnaire, and a question was run on the Okapi look motor based on that need.

The top 50 records recovered were clustered and clients were inquired to rank these Groups in request of preference.

They were then appeared individual records (titles, authors, source, date and abstract) and inquired to mark each record as applicable or non-relevant.

The exactness values of the Groups were then compared to the exactness values of the positioned records at tantamount edge levels. (If the Cluster positioned first by the client had 12 documents, it was compared with the exactness esteem of the positioned list at the top 12 records level. A comparable correlation was made for the records included in the first- and second-positioned clusters.)

As yield Clustering represented an overhead both for the framework (time and processing resources required to perform the clustering) and the client (time required to

assess the Cluster representations), it was essential to assess whether any benefits brought about by Clustering outweighed the accompanying overhead.

V. SIGNIFICANCE OF BEST EXACTNESS CLUSTERS

In some recent studies, analysts have concluded that clients are capable of recognizing the best (i.e. highest precision) clusters, and have based some of their execution correlations on this assumption. As seen from Table 4, our test discoveries do not support this assumption, and reveal that clients can't be relied on to recognize the best clusters. This raises questions about the validity of used best exactness Groups in surveying Clustering solutions. Moreover correlation of best exactness Groups against the positioned records has another flaw: it gives the Clustered yield an unfair advantage, as the following dialog indicates.

When the best exactness Groups were compared to the positioned records for each of the 55 client experiments, it was found that they obviously beated the positioned records in terms of exactness (Table 9).

Table 9 : Execution of best Groups versus positioned lists

Number of cases where Higher exactness given by:	Top Cluster level	Top 2 Groups level	Total
Best cluster(s)	33(60%)	39(71%)	72(65%)
Positioned lists	9(16%)	8(15%)	17(15%)
Equal n=55	13(24%)	8(15%)	21(19%)

However, this remarkable execution has little practical significance, since indeed randomly- made Groups are likely to outperform positioned records when sorted in exactness order. The reason is that the Cluster sizes are not extensive enough to have a dispersion of applicable records that converge to the normal figures, and discomparability from the normal produces both low- and high-exactness clusters. The smaller the Cluster sizes, the higher is the chance of outperforming the positioned list. This is because:

- The impact of discomparability is more pronounced: one extra applicable record makes a bigger distinction to exactness in a set of six records than in a set of 20 documents,
- Given a fixed number of documents, the more Groups there are, the more choices to select from.

To clarify this point, an test was performed to assess the degree to which Clustering formed Clusteres of records with higher exactness values than those could be anticipated under a arbitrary distribution. For each of the 55 queries, 100 arbitrary Cluster conveyances were created, with Cluster sizes matching those originally created. Exactness values were calculated for each of the Groups from these distributions, and the highest values were averaged to generate an estimated anticipated exactness for the best clusters. These values were then compared to the actual best exactness values accomplished in the experiments.

In 30 (55%) out of the 55 cases, the unique best Groups were beated by the normal best exactness esteem anticipated under arbitrary distribution. In the remaining 25 cases, the unique best Groups gave higher exactness values. However, although the distinction in terms of number of cases was in support of the arbitrary distribution, the unique best Groups had on normal 3% better exactness than the values anticipated under arbitrary distribution.

VI. PROPOSED SYSTEM

The proposed system introduced pattern mining based Topic Modelling techniques utilizing the Information filtering. The pattern mining based techniques have been used to utilize patterns to represent users' interest and have achieved some improvements in effectiveness, since patterns carry more semantic meaning than terms. Also, some data mining techniques have been developed to improve the quality of patterns (i.e. maximal patterns, closed patterns and master patterns) for removing the redundant and noisy patterns.

Advantages:

- The proposed approach is used to improve the accuracy of evaluating term weights.
- Because, the discovered patterns are more specific than whole documents.
- To avoiding the issues of phrase-based approach to using the pattern-based approach.

Problem Statement

It is proposed to model users' interest with multiple topics rather than a single topic under the assumption that users' information interests can be diverse.

We propose to integrate data mining techniques with statistical topic modelling techniques to generate a pattern-based topic model to represent documents and document collections. The proposed model MPBTM consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations representing the semantic meaning of each topic.

We propose a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. Patterns in each equivalence class have the same frequency and represent similar semantic meaning. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents.

We propose a new ranking method to determine the relevance of new documents based on the proposed model and, especially, the structured pattern-based topic representations. The Maximum matched patterns, which are the largest patterns in each equivalence class that exist in the incoming documents, are used to calculate the relevance of the incoming documents to the user's interest. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents.

IF systems obtain user information needs from 'user profiles'. IF systems are commonly personalized to support the long-term information needs of a particular user or a group of users with similar needs. In an IF process, the primary objective is to perform a mapping from a space of incoming documents to a space of user relevant documents. More precisely, denoting the space of incoming documents as D, the mapping rank : D -> R such that Rank(d) corresponds to the relevance of a document d. The filtering track in the TREC data collection was to measure the ability of IF systems to separate relevant from irrelevant documents.

The document filtering can be regarded as a classification task or a ranking task. Methods, such as Naive Bayes, kNN and SVM, assign binary decisions to documents (relevant or irrelevant) as a special type of classification. The relevance of a document can be modelled by various approaches that primarily include a term-based model, a pattern-based model, a probabilistic model and a language model.

Proposed Algorithm

A. Pattern Equivalence Class

Normally, the number of frequent patterns is considerably large and many of them are not necessarily useful. Several concise patterns have been proposed to represent useful patterns generated from a large dataset instead of frequent patterns such as maximal patterns and closed patterns. The number of these concise patterns is significantly smaller than the number of frequent patterns for a dataset. In particular, the closed pattern has drawn great attention due to its attractive features.

B. Definition

Definition 1 (Closed Itemset). *For a transactional dataset, an itemset X is a closed itemset if there exists no itemset X' such that (1) $X \subset X'$, (2) $supp(X) = supp(X')$.*

Definition 2 (Generator). *For a transactional dataset Γ, let X be a closed itemset and $T(X)$ consists of all transactions in Γ that contain X, then an itemset g is said to be a generator of X iff $g \subset X, T(g) = T(X)$ and $supp(X) = supp(g)$.*

Definition 3 (Equivalence Class). *For a transactional dataset Γ, let X be a closed itemset and $G(X)$ consist of all generators of X, then the equivalence class of X in Γ, denoted as $EC(X)$, is defined as $EC(X) = G(X) \cup \{X\}$.*

C. Algorithm Process

Algorithm *Document Filtering*

Input: user interest model $\mathbb{U}_E = \{\mathbb{E}(Z_1), \ldots, \mathbb{E}(Z_V)\}$, a list of incoming document D_{in}
Output: $rank_E(d), d \in D_{in}$

```
rank(d) := 0
for each d ∈ D_in do
    for each topic Z_j ∈ [Z_1, Z_V] do
        for each equivalence class EC_jk ∈ E(Z_j) do
            Scan EC_k,j and find maximum matched pattern MC^d_jk
            which exists in d
            update rank_E(d) using Equation (3):
                rank(d) := rank(d) + |MC^d_jk|^0.5 × f_jk × ϑ_D,j
        end for
    end for
end for
```

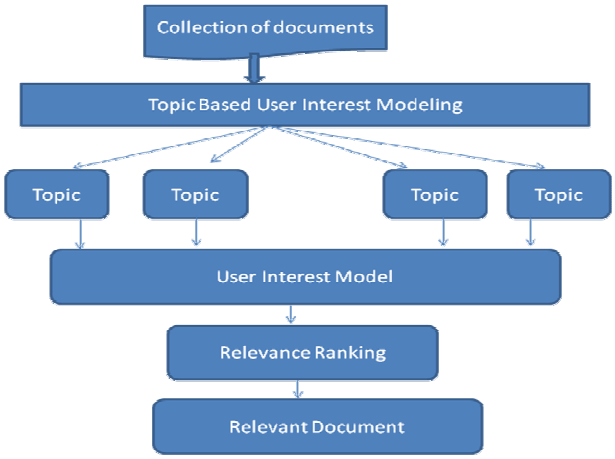

Fig.1 Overall Proposed Architecture

VII. CONCLUSION

This project presents an innovative pattern enhanced topic model for information filtering including user interest modeling and document relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interest's across multiple

topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modeling and the specificity as well as the statistical significance from the most representative patterns. The proposed model has been evaluated by using the RCV1 and TREC collections for the task of information filtering. In comparison with the state-of-the-art models, the proposed model demonstrates excellent strength on document modeling and relevance ranking.

Future Work

The proposed model automatically generates discriminative and semantic rich representations for modeling topics and documents by combining statistical topic modeling techniques and data mining techniques. The technique not only can be used for information filtering, but also can be applied to many content-based feature extraction and modeling tasks, such as information retrieval and recommendations.

REFERENCES

[1] Bastide Y., Taouil R., Pasquier N., Stumme G., and Lakhal L. (2000) "Mining frequent patterns with counting inference" ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75.

[2] Bayardo Jr R. J. (1998) "Efficiently mining long patterns from databases" in Proc. ACM Sigmod Record, vol. 27, no. 2, pp. 85–93.

[3] Beil F., Ester M., and Xu X. (2002) "Frequent term-based text clustering" Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discov. Data Minining, pp. 436–442.

[4] Cheng H., Yan X., Han J., and Hsu C.-W. (2007) "Discriminative frequent pattern analysis for effective classification" Proc. IEEE 23rd Int. Conf. Data Eng., pp. 716–725.

[5] Gao Y., Xu Y., Li Y., and Liu B. (2013) "A two-stage approach for generating topic models" in Advances in Knowledge Discovery and Data Mining, PADKDD'13. New York, NY, USA: Springer, pp. 221–232.

[6] Gao Y., Xu Y., and Li Y. (2013) "Pattern-based topic models for information filtering" in Proc. Int. Conf. Data Min. Workshop SENTIRE, pp. 921–928.

[7] Han J., Cheng H., Xin D., and Yan X. (2007) "Frequent pattern mining: Current status and future directions" Data Min. Knowl. Discov., vol. 15, no. 1, pp. 55–86.

[8] Robertson S., Zaragoza H., and Taylor M. (2004) 'Simple BM25 extension to multiple weighted fields' in Proc. 13th ACM Int. Conf. Inform. Knowl. Manag., pp. 42–49.

[9] Anuradha Awachar, Rajashree Bairagi, Vijayalaxmi Hegade and Mahadev Khandagale (2014), "An Overview of Ontology Based Text Document Clustering Algorithms", International Journal of Computer Sciences and Engineering, Volume-02, Issue-02, Page No (60-64

[10] Wang C. and Blei D. M. (2011) "Collaborative topic modeling for recommending scientific articles" in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 448–456.

[11] Xu Y., Li Y., and Shaw G. (2011), "Reliable representations for association rules" Data Knowl. Eng., vol. 70, no. 6, pp. 555–575.

[12] Zaki M. J. and Hsiao C.-J. (2002) "CHARM: An efficient algorithm for closed item set mining." in Proc. SDM, vol. 2, pp. 457–473.

