AN EFFICIENT METHODOLOGY FOR PATTERN-BASED TOPICS IDENTIFICATION IN DOCUMENT MODELING USING INFORMATION FILTERING

[1]Ms. U. Suriya, [2]Mr. B. Mohamed Apsal, [3]Mr. A. Mohanbabu

[1]Assistant Professor, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

[2]PG Student, II MCA, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

[3]PG Student, II MCA, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

Abstract:

Topic modeling has been widely accepted in the areas of machine learning and text mining, etc. It was proposed to generate statistical models to classify multiple topics in a collection of documents, and each topic is represented by a distribution of words. Although many variants of topic models have been proposed, most existing works are based on the bag-of-words representation that ignores the associations of words to represent topics. The word-based or term-based topic representations may not be able to semantically represent documents. Utilizing the proposed pattern-based topic model, users' interests can be modeled with multiple topics and each of which is represented with semantically rich patterns. In the proposed pattern-based topic model is adopted in the field of Information Filtering (IF). We proposed two novel models, Pattern-based Topic Model (PBTM) and Structural Pattern-based Topic Model (StPBTM). The main distinctive features of the proposed models include, user information needs are generated in terms of multiple topics; Document relevance ranking is determined based on topic distribution and topic related semantic patterns; Patterns are organized structurally based on the patterns' statistical and taxonomic features for representing user interests for each topic; significant matched patterns and maximum matched patterns are proposed based on the patterns' statistical and taxonomic features to enhance the pattern representations and document ranking. For information retrieval, we propose an unsupervised query expansion method, called Topical Pattern Query Expansion (TPQE), which expands a given query based on the topical patterns generated from the document collection by using the proposed pattern-based topic models. The results show that the proposed IF models significantly outperform state-of-the-art models and also prove the feasibility of the proposed query expansion model to deal with the short-query problem in IF.

Keywords: *- Topic Modelling, Pattern Model, Information, Filtering and Words.*

1. INTRODUCTION

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources [1], resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars. If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts. Alternatively, the data mining database could be a logical or a physical subset of a data warehouse.

Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts. Figures 1.1 and 1.2 illustrate this process.

As shown in Figure 1.1, almost all data enter the warehouse from the operational environment. The data are then "cleaned" and moved into the warehouse. The data continue to reside in the warehouse until they reach an age where one of three actions is taken [2]: the data are purged; the data, together with other information, are summarized; or the data are archived. An aging process inside the warehouse moves current data into old detail data.

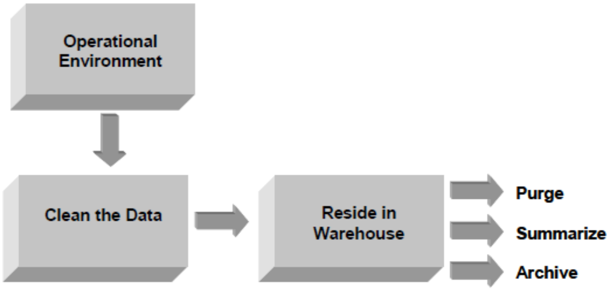


Fig 1.1: - Data Flow

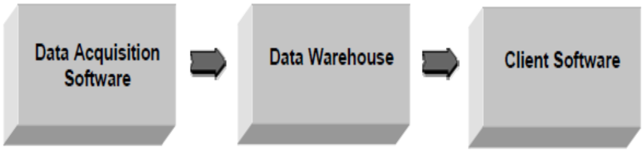


Fig 1.2: - Data Warehouse Architecture

Typically the data warehouse architecture has three components:

- Data acquisition software (back-end) which extracts data from legacy systems and external sources, consolidates and summarizes the data, and loads them into the data warehouse.
- The data warehouse itself contains the data and associated database software. It is often referred to as the "target database."
- The client (front-end) software, which allows users and applications (such as DSS and EIS) to access and analyze data in the warehouse.

These three components may reside on different platforms, or two or three of them may be on the same platform. Regardless of the platform combination, all three components are required.

There is some real benefit if your data is already part of a data warehouse. As we shall see later on, the problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. Furthermore, you will have already addressed many of the problems of data consolidation and put in place maintenance procedures [3]. The data mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then you will be better off with a separate data mining database.

1.1. Topic Modelling

Many data mining techniques have been used for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still a research issue. Traditional Information filtering models were developed using a term-based approach. The advantage of the term-based approach is its efficient computational performance. Term-based document representation tolerate from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based method have been used to utilize patterns to represent users' interest and have achieved some improvements in effectiveness [4, 5] since patterns carry more semantic meaning than terms. All these data mining and text mining techniques hold the assumption that the user's interest is only related to a single topic. However, in real world this is not necessarily the case. At any time, new topics may be introduced in the document, which means the user's interest can be diverse and changeable. Many mature term-based or pattern-based approaches have been used in the field of information filtering to generate users' information needs from a collection of documents. A fundamental assumption for these approaches is that the documents in the collection are all about one topic. However, in reality users' interests can be diverse and the documents in the collection often involve multiple topics.

Topic Modelling, such as Latent Dirichlet Allocation (LDA) [6] is a probabilistic model for collections of discrete data such as text collections. It can automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Two representative methods are Probabilistic Latent Semantic Analysis (PLSA) and LDA. However, there are two problems if we directly apply topic models for information filtering. The first problem is that the topic distribution itself is insufficient to represent documents due to its limited number of dimensions. The second problem is that the word based topic representation is limited to distinctively represent documents which have different semantic content since many words in the topic representation are repeated general words.

2. PROBLEM DEFINITION

The proposed topic model represents topics using patterns with structural characteristics which make it possible to interpret the topics with semantic meanings. As with existing topic models, the proposed model is application independent and can be applied to various domains. This research focuses on how the proposed pattern-based topic model can be used in the area of information filtering (IF) for constructing content-based user interest modelling, and additionally the research also investigates the feasibility of applying the pattern-based topic model to query expansion in information retrieval [7].

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' interests. The input data of IF is usually a collection of documents that a user is interested, which represent the user's long-term interests often called the user's profile. As mentioned before, users' information needs usually involve multiple topics. Hence, the proposed pattern-based topic modelling is applied to extract long-term user's interest through IF. Information retrieval (IR) typically seeks to find documents that are related to a user generated query from a given collection. The input data of IR is a query consisting of a number of terms which represent the user's short-term interest. One significant problem is that the length of queries is usually short and the keywords in a query are very often ambiguous or inconsistent. Topic modelling is an effective tool to manage large volumes of documents and user profiles [8]. As introduced in previous section, traditionally, the word-based topic representation is limited in its capacity to semantically represent documents and topics. And the word probabilistic distributions cannot identify those combinations of words that are more associated with each other within one topic. Many successful models used in IF and IR consider users' interests as single topics.

Nevertheless, in reality this is not the case. Users' profiles can include multiple topics and users' interests are diverse by nature. Some attempts to use multi-topic models have been studied in the fields of IF and IR, but most of them are restricted to using the topic distribution or bag-of-words features in topics for modelling users' interests and then ranking documents. With the purpose of contributing to the applications of IF and IR by using a novel topic modelling, we raise another research problem in this thesis.

3. EXISTING SYSTEM STUDY

The key component of an adaptive filtering system is the user profile used by the system to make the decision of whether to deliver a document to the user or not. In the early research work as well as some recent commercial filtering systems, a user profile is represented as Boolean logic [9]. With the growing computation power and the advance of research in the information retrieval community in the last 20 years, filtering systems have gone beyond simple Boolean queries and represent a user profile as a vector, a statistical distribution of words or something else. Much of the research on adaptive filtering is focused on learning a user profile from explicit user feedback on whether the user likes a document or not while interacting with the user. In general, there are two major approaches.

A typical classification system learns a classifier from a labeled training data set, and then classifies unlabeled testing documents into different classes. A popular approach is to treat filtering as a text classification task by defining two classes: relevant vs. non-relevant [10]. The filtering system learns a user profile as a classifier and delivers a document to the user if the classifier thinks it is relevant or the probability of relevance is high. The state of the art text classification algorithms, such as support vector machines (SVM), K nearest neighbors (KNN), neural networks, logistic regression and Winnow, have been used to solve this binary classification task.

Instead of minimizing classification error, an adaptive filtering system needs to optimize the standard evaluation measure, such as a user utility. Some machine learning approaches, such as logistic regression or neural networks, estimate the probability of relevance directly, which makes it easier to make the binary decision of whether to deliver a document. Many standard text classification algorithms do not work well for a new user, which usually means no or few training data points. Some new approaches have been developed for initialization. For example, researchers have found that retrieval techniques, such as Rocchio, work well at the early stage of filtering when the system has very few training data. Statistical text classification techniques, such as logistic regression, work well at the later stage of filtering when the system has accumulated enough training data. Techniques have been developed to combine different algorithms, and their results are promising [11]. Yet another example discussed in the following section is to initialize the profile of a new user based on training data from existing users.

It is worth mentioning that when adapting a text classification technique to the filtering task, one need to pay attention that the classes are extremely unbalanced, because most documents are not relevant. The fact that the training data are not sampled randomly is also a problem that has not been well studied.

4. PROPOSED SYSTEM METHODOLOGY

A fundamental assumption for these approaches is that the documents in the collection are all about one topic. However, in reality users' interests can be diverse and the documents in the collection often involve multiple topics. Topic modelling, such as Latent Dirichlet Allocation (LDA), was proposed to generate statistical models to represent multiple topics in a collection of documents, and this has been widely utilized in the fields of machine learning and information retrieval, etc. But its effectiveness in information filtering has not been so well explored. Patterns are always thought to be more discriminative than single terms for describing documents.

To deal with the above mentioned limitations and problems, in this paper, a novel information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is proposed. The main distinctive features of the proposed model include: (1) user information needs are generated in terms of multiple topics; (2) each topic is represented by patterns; (3) patterns are generated from topic models and are organized in terms of their statistical and taxonomic features; and (4) the most discriminative and representative patterns, called Maximum Matched Patterns, are proposed to estimate the document relevance to the user's information needs in order to filter out irrelevant documents.

In order to alleviate the ambiguity of the topic representations in LDA, in [12], we proposed a promising way to meaningfully represent topics by patterns rather than single words through combining topic models with pattern mining techniques. Specifically, the patterns are generated from the words in the word-based topic representations of a traditional topic model such as the LDA model. This ensures that the patterns can well represent the topics because these patterns are comprised of the words which are extracted by LDA based on sample occurrence and co-occurrence of the words in the documents. The pattern based topic model, which has been utilized in IF [13], can be considered as a ”post-LDA” model in the sense that the patterns are generated from the topic representations of the LDA model. Because patterns can represent more specific meanings than single words, the pattern-based topic models can be used to represent the semantic content of the user's documents more accurately compared with the word-based topic models. However, very often the number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent specific topics.

In this paper, we propose to select the most representative and discriminative patterns, which are called Maximum matched Patterns, to represent topics instead of using frequent patterns. A new topic model, called MPBTM is proposed for document representation and document relevance ranking. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents [14 ,15]. The original contributions of the proposed MPBTM to the field of IF can be described as follows:

1) Here, propose to model users' interest with multiple topics rather than a single topic under the assumption that users' information interests can be diverse.
2) And also 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.
3) The proposed 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.
4) Here, 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.

5. PATTERN BASED TOPIC MODEL

A two-stage approach is proposed to combine the statistical topic modelling technique with the classical data mining techniques, with the hope of improving the accuracy of topic modelling in large document collections. In stage one, the most recognized topic modelling method, Latent Dirichlet Allocation (LDA), is used to generate initial topic models. In stage two, the most popularly used term weighting method tf-idf and the frequent pattern mining method are used to derive more discriminative terms and patterns to represent topics of the collections [16]. Moreover, the frequent patterns reveal structural information about the associations between terms that make topics more understandable, semantically relevant and cover broader meanings.

5.1. Stage 1: - Topic Representation Generation

LDA is the typical statistical topic modelling and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents using the words that appear in the documents.

5.2. Stage 2:- Topic Representation Optimization

For most LDA based applications, the words with high probabilities in topics word distributions are usually chosen to represent topics.

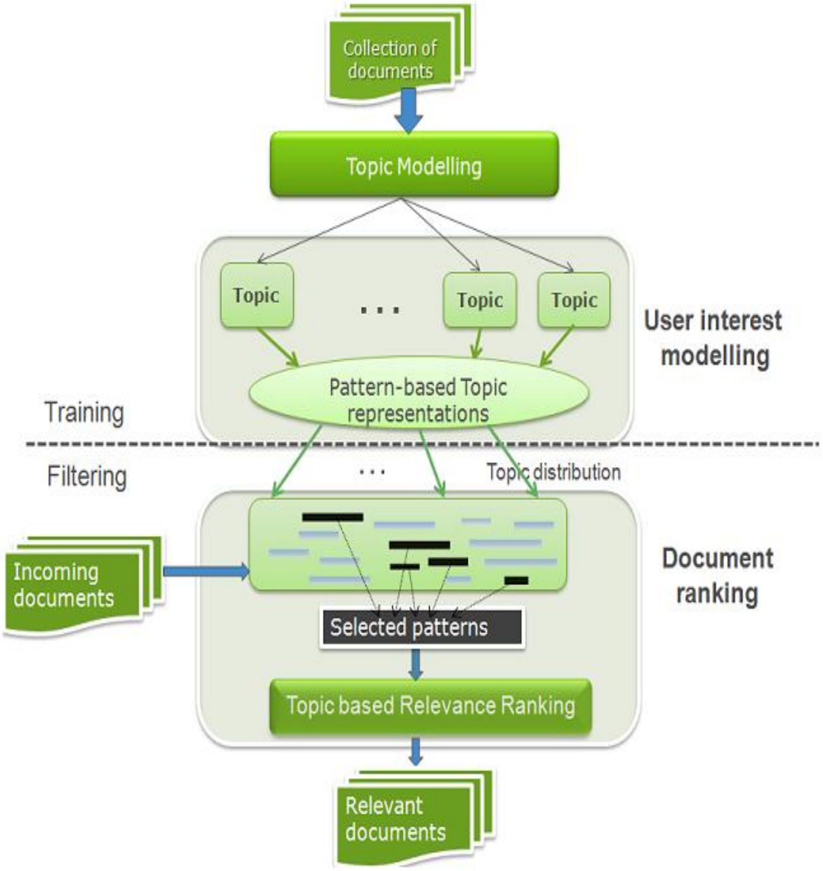
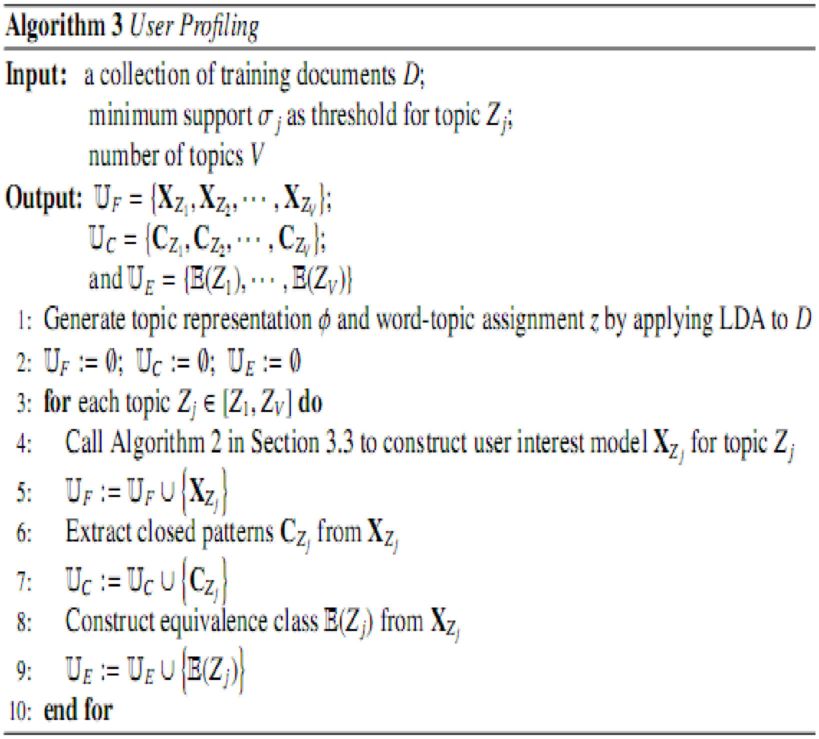


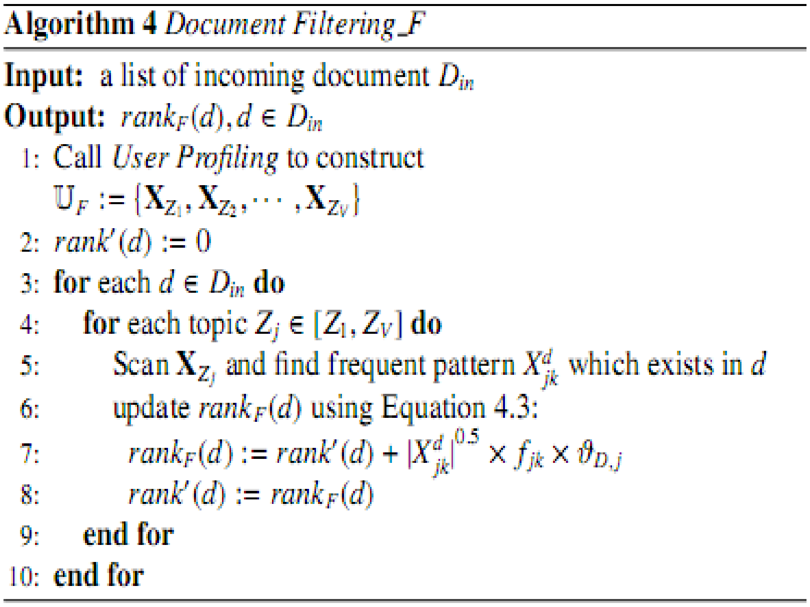
Fig 5.1: - The Structure of the Proposed IF Model

5.3. Algorithm for Pattern Model for IF

To understand the proposed models clearly, the algorithms of the proposed IF models (i.e., the PBTM (PBTM FP and PBTM FCP) models, the StPBTM (StPBTM SP and StPBTM MP) models) are given below. We divided this process into two algorithms: Algorithm User Profiling (i.e., generating user interest models) and Algorithm Document Filtering (i.e., relevance ranking of incoming documents). The former generates pattern-based topic representations to represent user's information needs. The latter ranks the incoming document based on the relevance of the discovered patterns. The algorithm Document Filtering actually is divided into four detailed algorithms, Document Filtering F, Document Filtering C, Document Filtering S, and Document Filtering M, for ranking incoming documents using the frequent patterns.



```
Algorithm 3 User Profiling
Input:  a collection of training documents $D$;
        minimum support $\sigma_j$ as threshold for topic $Z_j$;
        number of topics $V$
Output: $\mathbb{U}_F = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$;
        $\mathbb{U}_C = \{\mathbf{C}_{Z_1}, \mathbf{C}_{Z_2}, \cdots, \mathbf{C}_{Z_V}\}$;
        and $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$
Generate topic representation $\phi$ and word-topic assignment $z$ by applying LDA to $D$
$\mathbb{U}_F := \emptyset$;  $\mathbb{U}_C := \emptyset$;  $\mathbb{U}_E := \emptyset$
for each topic $Z_j \in [Z_1, Z_V]$ do
    Call Algorithm 2 in Section 3.3 to construct user interest model $\mathbf{X}_{Z_j}$ for topic $Z_j$
    $\mathbb{U}_F := \mathbb{U}_F \cup \{\mathbf{X}_{Z_j}\}$
    Extract closed patterns $\mathbf{C}_{Z_j}$ from $\mathbf{X}_{Z_j}$
    $\mathbb{U}_C := \mathbb{U}_C \cup \{\mathbf{C}_{Z_j}\}$
    Construct equivalence class $\mathbb{E}(Z_j)$ from $\mathbf{X}_{Z_j}$
    $\mathbb{U}_E := \mathbb{U}_E \cup \{\mathbb{E}(Z_j)\}$
end for
```



```
Algorithm 4 Document Filtering_F
Input:  a list of incoming document $D_{in}$
Output: $rank_F(d), d \in D_{in}$
Call User Profiling to construct
$\mathbb{U}_F := \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$
$rank'(d) := 0$
for each $d \in D_{in}$ do
    for each topic $Z_j \in [Z_1, Z_V]$ do
        Scan $\mathbf{X}_{Z_j}$ and find frequent pattern $X_{jk}^{d}$ which exists in $d$
        update $rank_F(d)$ using Equation 4.3:
            $rank_F(d) := rank'(d) + |X_{jk}^{d}|^{0.5} \times f_{jk} \times \vartheta_{D,j}$
            $rank'(d) := rank_F(d)$
    end for
end for
```

This part presents two innovative pattern-based topic models for information filtering including user interest modelling and document relevance ranking. The PBTM model and the StPBTM model generate pattern-based topic representations to model user's information interests across multiple topics. Since the number of all the patterns for topics is probably huge for modelling specific user's interests, all the topical patterns are partitioned into different groups in topics according to equivalence classes. Utilizing the hierarchical structure and partitions of topical patterns, the StPBTM model can more precisely model user information needs than the PBTM model.

In the filtering stage, the PBTM model chooses all frequent patterns or closed patterns; the StPBTM SP model selects partially significantly patterns to rank the relevance between user's interests and documents. In particular, the StPBTM MP model selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approaches incorporate the semantic structure from topic modelling and the specificity as well as the statistical significance from the representative patterns (frequent patterns closed patterns, significantly patterns, maximum matched patterns) with different kinds of strategies.

6. IMPLEMENTATION

High dimensional data are data characterized by few dozen to many thousands of dimensions. And any dataset represent able under a relational model is chosen as a High Dimensional Dataset [17]. According to that the following six different datasets were used, it is worth noting that the 20NG, Sports, Health, Society, and Local News.

Category	**No. of User Profiles**
20NG	412
Sports	300
Health	669
Society	442
Local News	254

Table 6.1: - The Category of Dataset

6.1. Performance Evaluation Parameters

As we can see from the experiment results, taking topics into consideration in generating user interest models and also in document relevance ranking can greatly improve the performance of information filtering [18]. The reason for the PBTM model and the StPBTM model achieving the excellent performance is mainly because we creatively incorporate pattern mining techniques into topic modelling to generate pattern based topic models which can represent user interest needs in terms of multiple topics.

Most importantly, the topics are represented by patterns which bring concrete and precise semantics to the user interest models. The outstanding performance of the StPBTM SP model over the PBTM FP model and the PBTM FCP model indicates the significant benefit of using the proposed significant patterns in estimating document relevance over using frequent patterns and frequent closed patterns. Moreover, the StPBTM MP model performs better than the StPBTM SP model, which is the best modelling among the entire pattern-based topic modelling for information filtering. This is because the proposed maximum matched patterns are most representative and quality patterns for modelling users' interests and relevance of documents.

All the proposed topic-based models outperform all the other baseline models including the pattern-based, phrase-based, and term-based models. As we have mentioned above, it is mainly because the topic based models represent the documents not only using patterns, phrases, or words, but also using topic distributions. Most importantly, the patterns, phrases or words used by the topic based models are topics related, which is a key difference from the pattern-based, phrase-based or word-based baseline models.

6.1.1. Recall and Precision

They are two well known measures of effectiveness in text mining. While Recall is a measure of correctly predicted documents by the system among the positive documents, Precision is a measure of correctly predicted documents by the system among all the predicted documents. The system is evaluated in terms of precision, recall and Fmeasure.

Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$precision = \frac{number\ of\ correct\ results}{number\ of\ all\ returned\ results}$$

$$recall = \frac{number\ of\ correct\ results}{total\ number\ of\ actual\ results}$$

6.1.2. F-Measure

F-measure combines precision and recall and is the harmonic mean of precision and recall.

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Several experiments were conducted with different query documents and the precision, recall and F-measure of the output was calculated.

6.1.3. Distortion

It is measured with the assistance of examination between Original dataset and changed dataset. Every tuple Xi in unique dataset [10], of m columns and each one column is of n characteristics , which is changed into Yi in altered dataset is utilized to register contortion in that tuple by ascertaining disparity between them through Euclidean separation by the equation.

6.1.4. Break Even Point

It is the point where recall equals precision. It is obtained by allowing the classifier to assign more categories. As a result, the recall increases and precision decreases until they become equal.

i) Recall

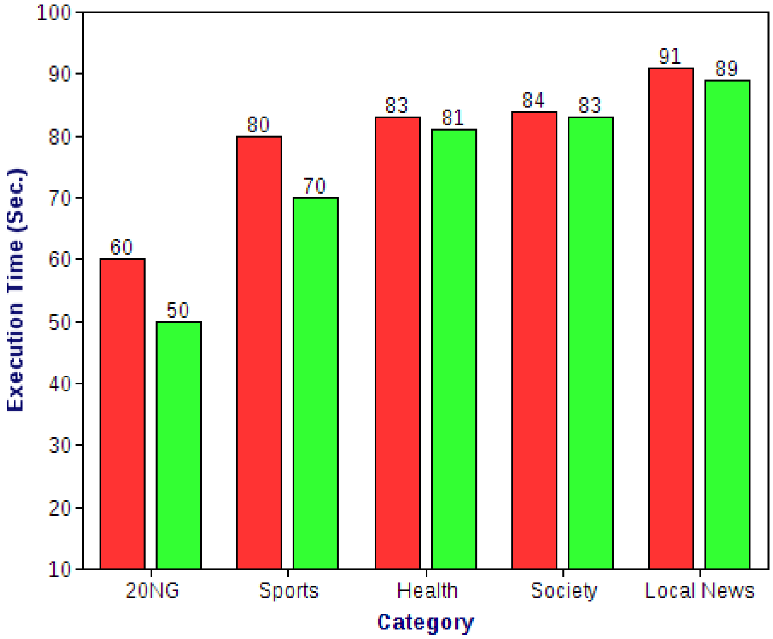


Fig 6.1: - Evaluation of Recall using Proposed Algorithm

ii) Precision

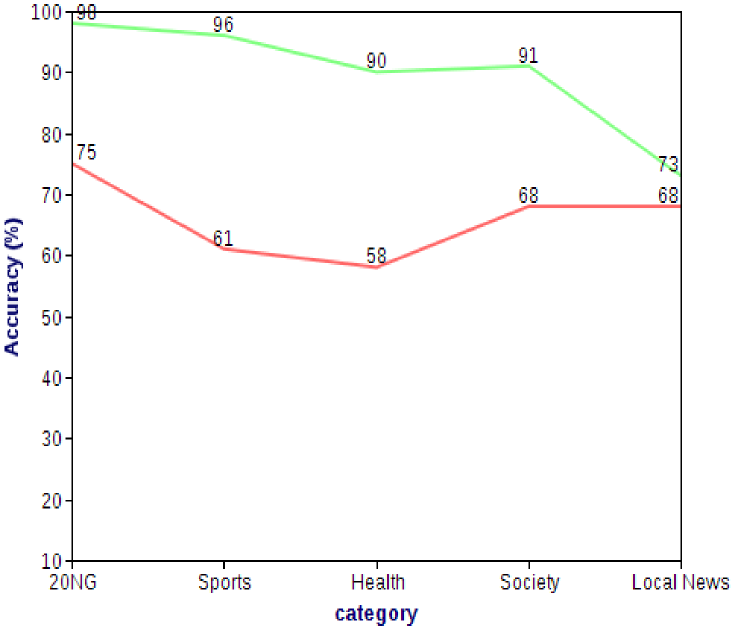


Fig 6.2: - Evaluation of Precision using Proposed Algorithm

iii) F-Measure

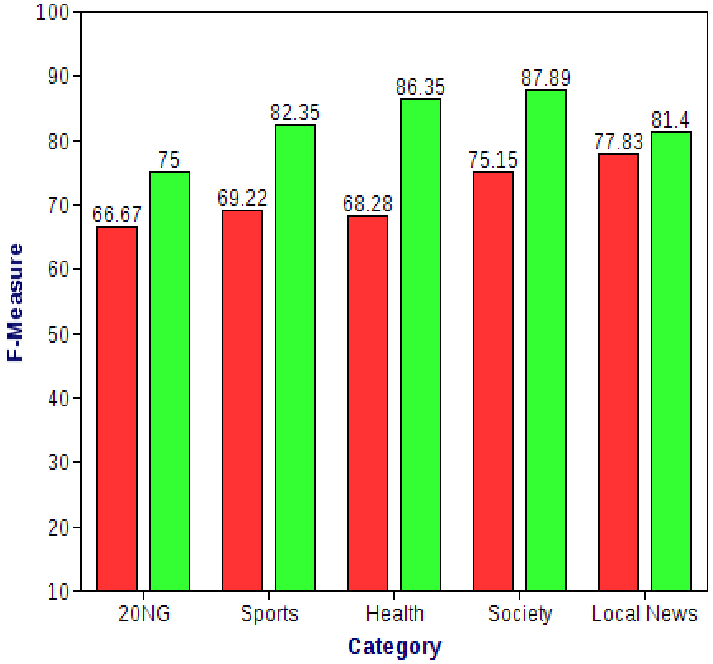


Fig 6.3: - Evaluation of F-Measure using Proposed Approach

7. CONCLUSION & FUTURE STUDY

This thesis starts by combating a long-standing challenge in web information, which is information overload. Understanding users' real information needs can help us distinguish most relevant information from large amounts of non-relevant information. We thus gave primary emphasis to seeking optimal models to accurately model underlying structure for users' interests. And utilizing the optimized user interest modelling to extract the relevance of documents and score the most relevant documents at top by developing relevance ranking system.

The user interest modelling in IF combines the statistical models with semantic feature representations, which outlines the user's interests with distribution of topics at a general level as well as interpretable features at detail level. On relevance ranking process, frequent patterns and closed patterns for the PBTM model, the proposed significantly matched patterns and maximum matched patterns for the StPBTM model, are selected to represent the relevance of documents. In Future, the information retrieval, the user provides a personally generated query for which it is difficult for the system to determine the particular topics involved. The proposed pattern based model is again adopted to determine user interested topics in the proposed TPQE model for query expansion. The potential relevant terms with the original query and related topics are discovered by utilizing the associations of words that are represented by topical patterns. The expanding terms are further refined afterwards by determining whether the query category is a focused query or a scattered query. For the document ranking, the TPQE estimates the relevance from aspects of determining query category at the general level, additionally analyzing expanded patterns with more specific features.

REFERENCES

[1] Gao, Y., Xu, Y., & Li, Y. (2014). Pattern-based topics for document modelling in information filtering. IEEE Transactions on Knowledge and Data Engineering, 27(6), 1629-1642.

[2] Zhong, N., Li, Y., & Wu, S. T. (2010). Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering, 24(1), 30-44.

[3] Li, Y., Zhou, X., Bruza, P., Xu, Y., & Lau, R. Y. (2008, October). A two-stage text mining model for information filtering. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 1023-1032).

[4] T.Kolda, D. O"Leary, "A semi-discrete matrix decomposition for latent semantic indexing in information retrieval". ACM Trans.Inform. Systems, vol. 16, pp. 322- 346 (1998).

[5] B. T. Bartell, G.W. Cottrell, R.K. Belew, "Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling". SIGIR, pp. 161-167 (1992).

[6] Geeganage, D. T. K., Xu, Y., & Li, Y. (2021). Semantic-based topic representation using frequent semantic patterns. Knowledge-Based Systems, 216, 106808.

[7] Zhang, L., Wu, Z., Bu, Z., Jiang, Y., & Cao, J. (2018). A pattern-based topic detection and analysis system on Chinese tweets. Journal of computational science, 28, 369-381.

[8] dos Santos, F. F., Domingues, M. A., Sundermann, C. V., de Carvalho, V. O., Moura, M. F., & Rezende, S. O. (2018). Latent association rule cluster based model to extract topics for classification and recommendation applications. Expert Systems with Applications, 112, 34-60.

[9] Nguyen, H., Xu, Y., & Li, Y. (2018). An ontology-based topic evaluation method for enhancing information filtering. In Data Science Meets Optimization Working Group (DSO Workshop).

[10] Bruzón, A. F., López-López, A., & Pagola, J. E. M. (2019). Improved Document Filtering by Multilevel Term Relations. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24 (pp. 85-95). Springer International Publishing.

[11] Papadakis, E., Gao, S., & Baryannis, G. (2019). Combining design patterns and topic modeling to discover regions that support particular functionality. ISPRS International Journal of Geo-Information, 8(9), 385.

[12] Li, C., Zhou, W., Ji, F., Duan, Y., & Chen, H. (2018, July). A deep relevance model for zero-shot document filtering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2300-2310).

[13] Na, L., Ming-xia, L., Hai-yang, Q., & Hao-long, S. (2021). A hybrid user-based collaborative filtering algorithm with topic model. Applied Intelligence, 51(11), 7946-7959.

[14] Bruzón, A. F., López-López, A., & Medina Pagola, J. E. (2018). Multi-level term analysis for profile learning in adaptive document filtering. Journal of Intelligent & Fuzzy Systems, 34(5), 3015-3026.

[15] JLiu, B., Li, C., Zhou, W., Ji, F., Duan, Y., & Chen, H. (2020). An attention-based deep relevance model for few-shot document filtering. ACM Transactions on Information Systems (TOIS), 39(1), 1-35.

[16] Banswal, D., Nagori, M., & Kshirsagar, V. (2018, July). Generating Homograph Models in Topic Modeling for Expediting User's Model Selection. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[17] Belwal, R. C., Rai, S., & Gupta, A. (2021). A new graph-based extractive text summarization using keywords or topic modeling. Journal of Ambient Intelligence and Humanized Computing, 12(10), 8975-8990.

[18] Wu, Y. (2018). Pattern-enhanced topic models and relevance models for multi-document summarisation (Doctoral dissertation, Queensland University of Technology).