Universität des Saarlandes
Naturwissenschaftlich-Technische Fakultät I
Fachrichtung Informatik

Masterarbeit

Personalization and Diversification of Detected Events in Social Media

vorgelegt von

Isha Khosla

am 30.05.2011

angefertigt unter der Leitung von

Dr.-Ing. Sebastian Michel

begutachtet von

Dr.-Ing. Sebastian Michel

PD Dr.-Ing. Ralf Schenkel

Declaration of Authorship

I, Isha Khosla, declare that this thesis titled, *Personalization and Diversification of Detected Events in Social Media* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Masters degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Event detection from an online stream populated by news articles, blogs, or tweets allows web users to stay up-to-date with the latest happenings or events. But the rapid growth of information related to events through online social media may overwhelm the personal interest of users. A user may not find any event of his interest in the Top-k retrieved events. This thesis is focussed on proposing a framework for presenting the users with the latest detected events matching their interests. Our contribution is summarized as follows: First, we propose multiple models based on classical tf*idf models and statistical language models for retrieving events matching to a user interest. Second, we refine the retrieved results by introducing user interest expansion technique based on Local Context Analysis. Third, to remove redundant events from the results and present the user with Top-k relevant, diverse, and novel events, we model this problem as *Maximum Diversity Quality Problem* and *Maximum Min-Diversity Quality Problem*. These problems are formulated as quadratic integer program. We propose greedy and linear programming solution for these problems. We present experimental studies based on the New York Times Archive validating both the quality of retrieval models and performance of greedy solution in comparison to linear programming solver for Maximum Diversity Quality Problem.

Acknowledgements

I would like to thank my advisor Dr.-Ing. Sebastian Michel for giving me the opportunity to pursue this thesis under his supervision and his timely and valuable inputs. I am grateful to him for bringing an interesting and challenging research topic to my attention. I also thank him for his persistent support and patience through many discussions we had through the course of the thesis. He has been an excellent advisor. I thank my parents for being a source of continued emotional support and teaching me to aim high. I am thankful to all my friends, specially Megha and Sharat for their support and encouragement.

Contents

List of Figures

List of Tables

Thesis Organization

The thesis is organized as follows:

Chapter 1 - Introduction : This chapter provides the introductory material, stating the problem and the solution briefly.

Chapter 2 - Background : This chapter covers the basics required to understand the algorithms and techniques used for solving the problem. It provides the basic knowledge related to language models, TF*IDF, query expansion techniques and other information retrieval concepts relevant to our work.

Chapter 3 - Personalized Retrieval of Detected Events : This chapter focuses on one part of our problem of personalizing and diversifying the detected events. The work in this chapter is focussed on proposing retrieval models for personalizing the detected events.

Chapter 4 - Refining Personalized Events using Local Context Analysis : This chapter explores user interest expansion techniques related to Local Context Analysis in order to improve the retreived results.

Chapter 5 - Diversity in Personalized Events : This chapter aims to propose methods for bringing novelty and diversity in the personalized results. Some methodologies are explored and developed to handle the issue of diversity for improving the personalized results.

Chapter 6 - Diversity measures : This chapter discuss the formulations of diversity measures in detail.

Chapter 7 - Experiments and Evaluation : This chapter discuss the experimental work related to the validation of quality of retrieval models and performance of greedy methods.

Chapter 8- Conclusion : The thesis work in concluded in this chapter.

At last, we have included Appendices to see the results obtained by experiments and Bibliography stating the refered material.

Chapter 1

Introduction

With the advent of systems which detect events, users will be automatically informed about the latest events at the earliest. The problem, however, is that these systems broadcast the same detected events to all users, ignoring their personal interests. For instance, consider a system which presents the latest events by monitoring the streams of news documents, tweets, or blogs. Further, consider some major events which are going to happen in the upcoming week - related to cricket: *cricket worldcup finals*, the others related to technology or gadgets: *launch of ipad2, launch of android phone*, or related to entertainment: *Eurovision Song Contest*, have extensively populated the stream. At the another end, we have three users of our system, one is always interested to know about latest electronic gadgets, the other user has a lot of interest in following musical concerts and the last user is a follower of cricket. A system without the component of personalization will return all the users with the same top-k detected events as shown in Figure 1.1.

Unfortunately, the interaction of users with such systems can be reviewed as an impersonal affair. As it might be possible that the user do not come across any of the events

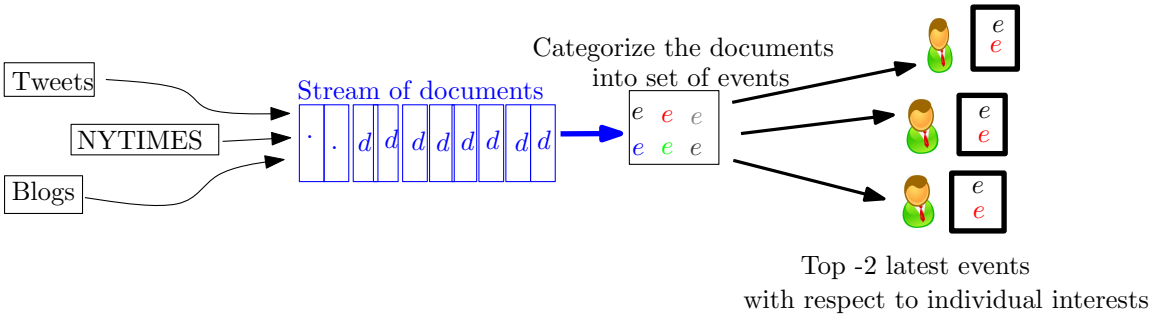


FIGURE 1.1: Unpersonalized Top-2 events returned by the system

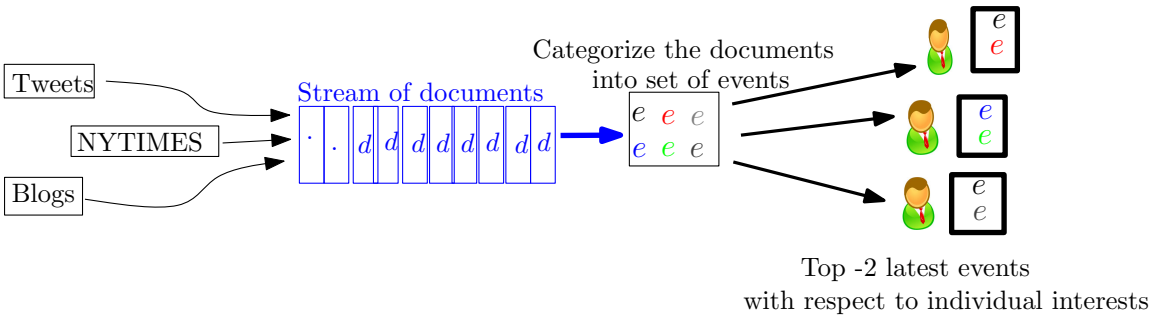


FIGURE 1.2: Personalized Top-2 events returned by the system

of his interest in the top-k detected events retrieved by the system. From Figure 1.1 we observe that all the users are presented with the same top-k results, which will bring dissatisfaction in the users. This limitation has motivated us to incorporate personalization component into the work on detecting events. Moreover, we want to incorporate diversity in the personalized results in order to present the user with non-redundant results of his interest. We propose a framework, called *The Personalization and Diversification*, which will present the user with top-k relevant, diverse, and novel events according to his interests (refer Figure 1.2). Our framework achieves the following goals.

1. Retrieves the relevant latest events according to a user interest, and incorporates flexibility in the system by providing the freedom to users to specify their interests as keywords.

2. Refine the ranking of retrieved events and eliminate the redundant events to minimize the dissatisfaction among users.

1.1 The Personalization and Diversification Framework

In order to achieve the above stated goals, our framework is an integration of three modules, i.e, Retrieval, Expansion and Diversification (refer Figure 1.3). Their functionalities are briefly described below,

- **R**etrieval - This is the basic module that retrieves x ($x >> k$) number of relevant events according to a specified user interest. In our work on retrieving relevant detected events, we propose four different models - Event as a Huge Document-Language Model (EHD-LM), Event as a Set of Documents-Language Model (ESD-LM), Event as a Huge Document-TFIDF (EHD-TFIDF), and Event as a Set of

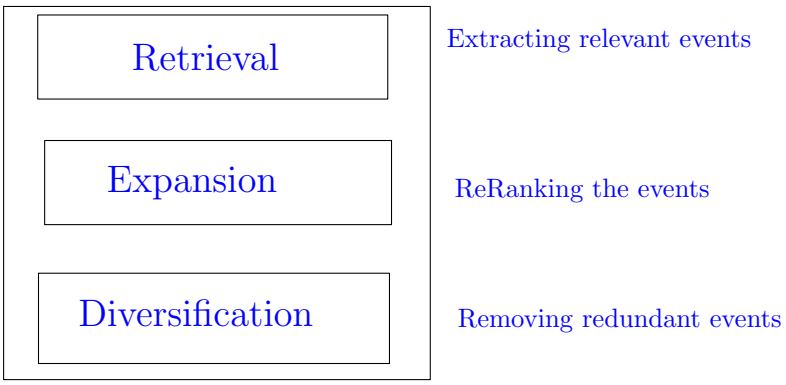


FIGURE 1.3: The Personalization and Diversification Framework

Documents -TFIDF (ESD-TFIDF). These models are designed using baseline measures as language models [1] and TFIDF [2]. Further details related to these models are provided in Chapter 3.

- **E**xpansion - The relevance of retrieved results is directly affected by the specificity maintained by a user in framing their interests as keywords. Sometimes, the inability of users in specifying their interests as informative and well defined may result in less relevant topics to be ranked higher. Actually such cases will be quite common in our problem setting, for example, a user who is interested in knowing about topics related to WorlCup of Cricket, in general, may mention his interest as "Sports" rather as "Cricket" or "WorldCup Cricket". And the retrieval module returns the detected events related to various sports (not only topics related to cricket). This problem is addressed by the expansion module by refining the ranking of the retrieved results. This module supports user interest driven expansion technique. Our work on modeling the expansion is related to the work on [3], [4]. The key idea of this technique is to first measure the co-occurrence of popular tags and terms which are linked to the documents in the topics retrieved by the Retrieval module. Then they append those tags and terms to the initial user interest which have the highest co-occurrence measure. At last, results are re-ranked according to the modified user interest. The precise details related to the evaluation of tags and terms as popular and the formulation of co-occurrence measure are described in Chapter 4.

- **D**iversification - There has been substantial research on the problem of diversity in the web search results [5]. Moreover, the problem of diversity has been addressed in the context of personalization and recommendations system, too [6]. But to address the issue of diversity in the personalization of detected events is novel. The main motivation of our work on diversification is to decrease the similarity

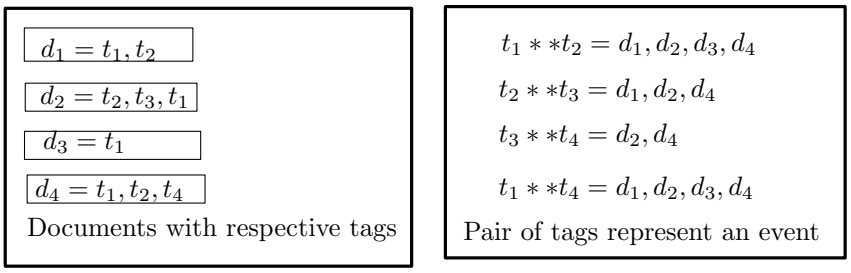


FIGURE 1.4: Event Representation as a Tag Pair

among the retrieved events to increase the satisfaction in user for the presented results (detected events). For example, suppose there is a upcoming event in a week for the technophile, e.g., launch of ipad2. On the other end, a user who is not very much interested in the launch of ipad2 but more interested in knowing about Apple products in general, specified his interest as " Apple Products" to retrieve the relevant events. But as the incoming sources (blogs, tweets or news) of information are extensively populated by discussion or articles on ipad2, it is obvious that the user is presented with top-k detected events which all are related to ipad2 features, price and release date and a user may not get any event related to other apple products. Though the events are relevant, they contain repetitive information and there is no variety in the results. To address this issue we eliminate similar events from the retrieved events in order to obtain top-k distinct and relevant events. In our work, we have proposed multiple diversity measures and algorithms to remove redundancy in the results, refer chapters 5 and 6.

Moreover, in our proposed framework of personalizing and diversifying the detected events, we represent an event as a pair of tags. Consider an incoming stream of documents d_1, d_2, d_3 and d_4 where each document is annotated with tags as shown in the Figure 1.4. Like, for documents d_2 and d_4, one possible event can be $t_3 * *t_4$. The other possible events for the above set of documents are given in the Figure 1.4. Then we personalize the detected events according to on the user's interest. We define user interests as keywords like *politics, education, fashion, economic crises* associated with each user. These keywords can be obtained either actively by asking the user or passively by storing user's application usage history (e.g., clicks [7], user profiling [8]).

Overall, in our proposed work on *Personalization and Diversification of Detected Events in Social Media*, we aim to provide a solution to personalize events detected from an online stream populated by news articles, tweets,or blogs according to user information needs. Moreover, we also aim to remove redundancy in the retrieved personalized results,

as we believe that removing redundant events will bring variety and freshness in the personalized results.

Chapter 2

Background

Our work on personalizing emerging topics and introducing novelty and diversity in the personalized results touches various research topics such as language models, local and global query expansion techniques, linear programming, and greedy solutions to NP hard problems. In the following, we briefly review all of these topics.

2.1 Language Models

In the information retrieval community, Ponte and Croft [9], [10] were the first to introduce the language modelling approach. Their work proposes a new way to score a document for a given query, also called query likelihood scoring model. Intuitively, given a query Q, the query likelihood retrieval model checks that whether the user would prefer or like to see the document D for a query Q. The score of document D w.r.t query Q is defined as the conditional probability $p(Q|\theta_D)$. The θ_D can be any language model.

$$score(Q, D) = p(Q|\theta_D)$$

However, the choice of a specific model can affect retrieval performance significantly. In our work, we adapted a unigram multinominal language model for θ_D [11]. Moreover, we incorporated Jelinek Mercer Smoothing model [12] to improve the accuracy of the estimated language model in general. However, the other models in the literature are multiple Bernoulli [9] and Poisson model [13].

These three models (multinominal, multiple bernoulli and multiple poisson) make separate assumptions about term or word occurrences in a document. The multinominal model assumes that every word occurrence and multiple occurrences of the same word is independent. The multiple bernoulli model is a binary model, it makes the assumption

that only occurrences of different words are independent. The multiple poisson model also assumes that the occurrences of different words are independent but it is not a binary model and captures word frequencies. However, much of the research work is carried out on the multinominal model and there has been some evidences that multinominal outperforms multiple bernoulli in the work of retrieving subtopics [1]. Whereas poisson model appears to more promising [13] but there has not been much work on this model.

2.2 TFIDF Model

Conceptually, the TFIDF [14], [2] statistical measure is the simplest retrieval/ranking measure which evaluates how relevant a document is to a query. The relevance increases proportionally to the number of times a query terms appears in a document but is offset by the frequency of the word in the collection. In TFIDF model, a weight which can be called as $tf * idf$ weight is assigned to every query term t for each document d. A $tf * idf_{t,d}$ assigns to term t, a weight in document d that is large when t occurs many times within a small number of documents; lower when the term occurs fewer times in a document, or occurs in many documents (means very less relevant); lowest when the term occurs in virtually all documents. Hence the score of query q for a document d is given as

$$score(q, d) = \sum_{t \in q} tf * idf_{t,d}.$$

In our problem setting, we extend this measure to evaluate the relevance of an event (in a collection of documents). They are discussed in detail in Chapter 4.

2.3 Local and Global Query Expansion Techniques

Basically, the research on query expansion techniques is splitted into two techniques (global and local).

2.3.1 Global Query Expansion Techniques

Global expansion techniques reformulate query terms independently of the query and query results. Moreover, they assume that words related in a corpus has high probability of cooccurring in the documents of that corpus. Term clustering [15], [16] is one of the

earliest research work on global techniques. It generates clusters of terms that cooccurred together in the corpus and later use those clusters for expansion. But if the query term has multiple meanings, term clustering will add terms related to multiple meanings and make the query more ambiguous. Hence the retrieved results would not be very effective or relevant. The recent work on global query expansion methods is more promising and can be classified into

1. Query expansion with Word Net [17].
2. Query expansion by automatic thesaurus generation [18].

2.3.2 Local Query Expansion Techniques

The local techniques reformulate a query w.r.t the documents that intially appeared to match the query. The idea of using the relevant or top ranked documents for query expansion was introduced by Attar & Fraenkel [19]. They incorporated this idea for term clustering [15] and then use the clusters for query expansion. Later, another local technique was proposed based on relevance feedback [20]. The idea behind relevance feedback is to take the results that are initially returned from a given query and to use information about whether or not those results are relevant to perform a new query. The relevance feedback can be categorized into: explicit feedback and implicit feedback.

The explicit feedback is directly given by the users of a system. Users grade the results and their grading will suggest whether the retrieved result is relevant or not. The relevance feedback information needs to be integrated with the original query to improve retrieval performance, such as the well-known Rocchio Algorithm [11]. But in the implicit relevance feedback, the user behavior is studied to gather their choice of considering document relevant or irrelevant. It can done by tracking the users click history, by tracking his profile and other page browsing or scrolling actions [8]. In the implicit relevance feedback, the user may or may not be informed that their behavior is studied for judging the relevance of documents.

Both of these techniques have shortcoming. The drawback in global techniques is that it requires analyses of the whole corpus and the major drawback in local techniques is that they require explicit feedback or judgements. To overcome these drawbacks, we introduce a new technique (known as local context analysis [3], [4], [15]) which is a combination of local and global techniques. In local context analysis techniques, we assume that the good query expansion terms tend to co-occur with all the query terms in the top ranked documents.

2.4 Linear Programming

We used the linear programming [21], [22], in our work to design models for refining the personalized results. We refine the personalized results to present variety and novelty in the personalized results. The variety or novelty in the results is attained by eliminating the duplicate results from the top-k ranking. We addressed the problem of redundancy in the results as the diversity problem. In our work to address the problem of diversity, we formulated the problem into mathematical models, such as linear integer program.

Formally, a linear programming problem is defined as maximizing or minimizing a linear function subject to linear constraints. It can also be stated as a mathematical model for determining a way to achieve the best outcome in a given mathematical model for a given objective function and a list of constraints. Linear programming models are applied in modeling variety of problems such as planning, routing, scheduling, and assigning.

In our work, we adapt linear integer program to solve the quadratic integer program formulation of diversity problem. Further details related to the modeling of the diversity problem are discussed in the Chapter 5. In order to obtain computational estimates, LP solvers are used to solve linear programming problems. In the market, various standard LP solvers are available, such as, CPLEX, glpk, etc. In our work, we evaluated our mathematical models on *CPLEX LP Solver*.

CPLEX is a powerful and stable Linear Programming (LP), Mixed-Integer Programming (MIP), Quadratically Constraint Programming (QCP) and Mixed-Integer Quadratically Constraint Programming (MIQCP) solver based on the Cplex Callable Library from *IBM ILOG CPLEX* [23]. We used a modeling layer called CONCERT in JAVA in the CPLEX Optimizer to solve the linear integer program. The reason for us to choose CPLEX solver is that the latest version, i.e., CPLEX 12.2, is considered to be on average 50% faster than the earlier versions. Its use is free for academic purposes.

Though the attainment of exact solutions for the diversity problem rely on the methods such as branch-and-bound (which can be implemented using solvers), solving a large sized problem with an exact method takes an excessively long time. Therefore the use of LP solvers is not recommended for realistic problems, like, retrieval and ranking problems in search engines. Hence, for solving problems in real time setting we prefer to sacrifice optimality to obtain fairly good solutions within an acceptable time by means of heuristics. In our work, we proposed greedy heuristics as an alternative solution. The greedy solution guarantees to provide a feasible solution in a negligible time.

2.5 Greedy Heuristics

The greedy heuristics can be used to achieve effective solutions to the problems such as coverage, scheduling, assigning, and planning. The greedy Heuristic we use is a stepwise and iterative procedure. A cost function or objective is defined and goal can be either to maximize or to minimize the cost. At each step, it selects the unit which maximizes or minimize the cost. The nature of greedy algorithms may easily yield feasible solutions to the problem. Moreover, it has the advantage of being fast and producing reasonably efficient solutions. There is tremendous literature on the use of greedy algorithms in various problem settings, like, greedy approximation algorithm for set cover problems [24], parallel greedy solution to addressing the MaxCover problem in large scale commodity clusters [25], and so on.

In this chapter, we have provided the preliminary introduction to the concepts and methodologies related to our work. Briefly, in Chapter 3, we refer the work on language models and TFIDF; in Chapter 4, we refer the work on query expansion techniques in designing the models related to our work; and in chapters 5 and 6, we refer to the work on linear programming and greedy heuristics.

Chapter 3

Personalized Retrieval of Events

The retrieval process starts with real-time extraction of the *events* as per user's interest from a stream of documents, tweets, or blogs. An event will be detected and presented to a user if the event related data (documents, tweets or blogs) in the incoming stream is frequently observed in the specified time interval. Our focus here is to *personalize the detected events*. But why do we need to personalize the detected events? For illustration, consider some major events which are going to happen in the upcoming week - one related to sports: *cricket worldcup finals*, the others related to technology or gadgets: *launch of ipad2, launch of android phone*, or related to entertainment: *Oscar awards celebration*, have extensively populated the stream. There will be many events retrieved due to number of events happening around, may be none of user's interest. Hence, emerges the need to retrieve events satisfying the user information needs. Like, in this case, a technophile would be excited to know the *launch of ipad2*, a cricket fan would be interested to know *cricket related topics*, and a theater follower would prefer to know about *best actor nominees for the Oscar awards* rather about other events.

Our proposed model on personalization organizes retrieval results (events) based on the user's interest. We represent user interests as keywords like *politics, education, fashion, economic crises* associated with each user. These keywords can be obtained either actively by asking the user or passively by storing user's application usage history (e.g., clicks [7], user profiling [8]).

In the following, we discuss several retrieval approaches based on the modelling of events. One possible modeling approach is to consider an event as a cluster of documents and the other is to represent an event as one huge document formed by concatenating its member documents. The former modeling approach is termed as *Event as a Set of Documents (ESD)* and the later is termed as *Event as a Huge Document(EHD)*. For illustration, refer Figure 3.1 where an event e contains a set of documents a, b, c, d. Further, each

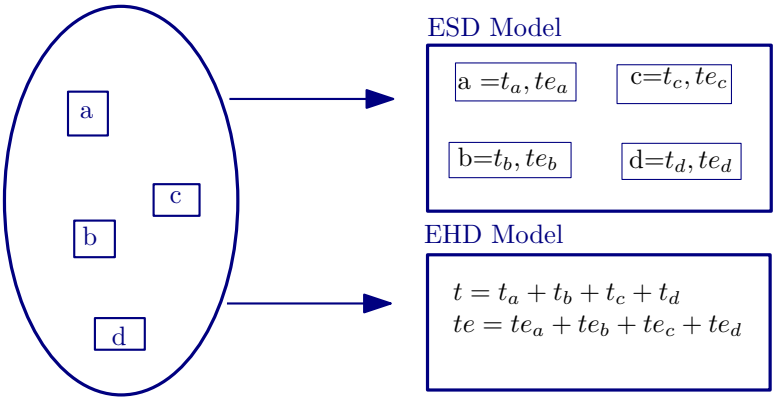


FIGURE 3.1: Representation of Event Models - ESD and EHD

document, say a, contains a set of tags t_a and a set of terms te_a. In the ESD model representation, an event is considered as a set of documents and each document contains a set of terms and a set of tags. Whereas in the EHD model, an event is considered as a huge document with tags and terms set formed by concatenating the terms and tags of its member documents.

3.1 Retrieval Approaches

We studied TFIDF and language models for retrieving personalized emerging topics from the incoming stream of documents. Each of these retrieval approaches is defined for both of our event models, i.e., *Event as a Set of Documents* and *Event as a Huge Document*. Our work on language models for retrieving events is related to the research work done on building language models for *ranking/retrieving clusters* [26][27]. The TFIDF work is extended from the traditional TFIDF [14] statistical measure. It is the simplest retrieval/ranking measure which evaluates how relevant an event is to user interest by computing term frequencies of words. The relevance increases proportionally to the number of times a word appears in the document (ESD) or events (EHD) but is offset by the frequency of the word in the collection. For a given user interest u, we rank events on the basis of quality measure Q_m, obtained through the respective retrieval models.

3.1.1 Language Model Based on EHD

In the Event as a Huge Document model (EHD), the quality measure Q_m is defined for a user interest u and an event e as the logarithm of the probability $P(u|e)$ of user

interest u being generated by the terms and tags of an event e. It is obtained as

$$Q_m(u, e) = \log P(u|e). \tag{3.1}$$

And for an event e, the maximum likelihood probability of user interest u, i.e., $P(u|e)$ is computed using language models. The language model based on EHD is extended from one of the cluster retrieval language model discussed in [26]. Their work is related to retrieving/ranking clusters based on the likelihood of generating the query Q, i.e. $P(Q|Cluster)$. The documents in the same cluster are concatenated and each cluster is considered as a huge document. $P(Q|Cluster)$ is specified by the cluster language model as

$$P(Q|Cluster) = \lambda P_{ML}(Q|Cluster) + (1 - \lambda)P_{ML}(Q|Collection)$$

where $P_{ML}(Q|Cluster)$ and $P_{ML}(Q|Collection)$ are maximum likelihood language models. Similarly in the EHD model, we build language models and apply smoothing on the whole event, and then retrieve events based on the likelihood of generating the user interest. Hence the cluster retrieval language model [26] acts as a baseline model for building *language model based on EHD*.

We use the following notation in designing our EHD language models.

d	document
e	event
T_e	Bag of tags for an event
T_{e_e}	Bag of terms for an event
T	Bag of tags for a collection of events
t_e	Bag of terms for collection of events
λ, β	Smoothing parameters
$\{w_1, w_2, ..., w_m\}$	words in a user interest

TABLE 3.1: Notation - EHD Model

In order to measure the quality of an event, i.e., Q_m, we need to make a probabilistic estimation or likelihood of user interest u being generated by an event e. The $P(u|e)$ is defined using language models as these models are effective in capturing the characterstics of events. We now discuss the computation of $P(u|e)$ using language models.

We start by applying an unigram language model to $P(u|e)$. It states that keywords $(w_1, w_2, ..., w_m)$, in a user interest u, are probabilistically independent of each other for an event e such that,

$$P(u|e) = P(w_1|e)P(w_2|e)...P(w_m|e) = \prod_{i=1}^{m} P(w_i|e)$$

$$or,\ \log P(u|e) = \sum_{i=1}^{m} \log P(w_i|e). \tag{3.2}$$

Thereby, we reduce the problem to that of estimating probabilities of individual words in the user interest. Thereafter, $P(w_i|e)$ is be defined as the maximum likelihood event language model M_e based on the count of words in an event e. But the problem with this event language model M_e is that an unseen word in the event e would get zero probability, making all queries containing an unseen word have zero probability for the entire user interest u and an event e. This is clearly undesirable. To solve this problem, *we smooth our event language model using a collection language model.* And the collection language model M_c are defined as maximum likelihood language models based on word frequencies in the collection as a whole. Hence, we obtain $P(w_i|e)$ as

$$P(w_i|e) = (1-\lambda)M_e(w_i) + (\lambda)M_c(w_i). \tag{3.3}$$

$\lambda \in (0,1)$ controls the relative weight given to the event language model M_e and the collection language model M_c (similar to the Jelinek-Mercer smoothing model [12][14]). Further, an event comprises of tags and terms collected from its member documents. And the likelihood of w_i as a tag and a term in an e is different. Hence to incorporate term and tag importance in an event language model M_e and collection language model M_c, we define them as

1. M_e is defined as a linear combination of $M_{et}(w_i)$ and $M_{et_e}(w_i)$, which are maximum likelihood language models based on the count of w_i appearing in an event as a tag and a term respectively.

   $$M_{et}(w_i) = P(w_i|T_e) = \frac{|\{w_i|w_i \in T_e\}|}{|T_e|},$$

   where T_e is a bag of tags for an event e.

   $$M_{et_e}(w_i) = P(w_i|T_{e_e}) = \frac{|\{w_i|w_i \in T_{e_e}\}|}{|T_{e_e}|}.$$

   where T_{e_e} is a bag of terms for an event e.

2. M_c is defined as a linear combination of $M_{ct}(w_i)$ and $M_{ct_e}(w_i)$, which are maximum likelihood language models based on the count of w_i appearing in the collection of events as a tag and a term respectively.

   $$M_{ct}(w_i) = P(w_i|T) = \frac{|\{w_i|w_i \in T\}|}{|T|},$$

where T is the bag of tags obtained from the collection of events.

$$M_{ct_e}(w_i) = P(w_i|T_e) = \frac{|\{w_i|w_i \in T_e\}|}{|T_e|}.$$

where T_e is the bag of terms obtained from the collection of events.

Finally we obtain $M_d(w_i)$ and $M_c(w_i)$ as

$$M_e(w_i) = (1-\beta)M_{et}(w_i) + \beta M_{et_e}(w_i), \text{ and} \tag{3.4}$$

$$M_c(w_i) = (1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i). \tag{3.5}$$

where $\beta \in [0,1]$ controls the relative weight given to the tag and term models.

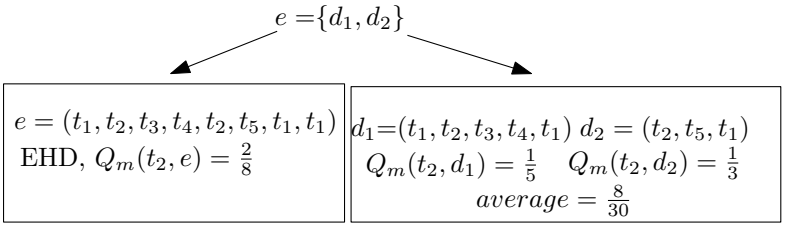
And from equations (3.9), (3.4) and (3.5), the probability of a word w_i in a user interest u of being generated by event e is given as

$$\begin{aligned} P(w_i|e) = ((1-\lambda)((1-\beta)M_{et}(w_i) + \beta M_{et_e}(w_i))) \\ +(\lambda)((1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i))). \end{aligned} \tag{3.6}$$

Hence, we obtain Q_m from equations (3.1) and (3.6) as

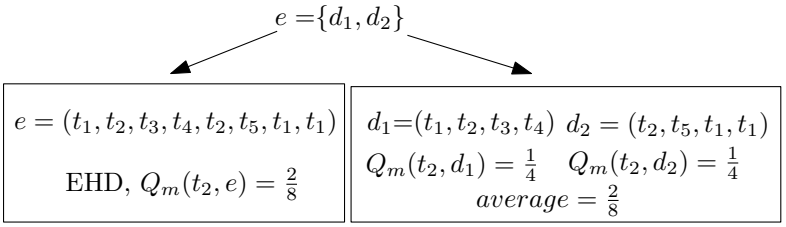
$$\begin{aligned} \mathbf{Q_m(u,e)} = \sum_{i=1}^{m} \{\log((1-\lambda)(\beta M_{et}(w_i) + (1-\beta)M_{et_e}(w_i))) \\ +(\lambda)(\beta M_{ct}(w_i) + (1-\beta)M_{ct_e}(w_i)))\}. \end{aligned} \tag{3.7}$$

The derivation of above Q_m is independent of the representation of the documents in an event. Though it makes it simple yet less powerful (intuitively) for an event which have hetrogenous documents, as we smooth out the effects or differences the representations of individual documents will make to the quality measure. On the other end, the Q_m can be effective or precise for an event which has a homogeneous collection of documents. For example, suppose there is an event e and it contains 2 documents, d_1 and d_2. The document d_1 consists of a bag of terms (t_1, t_2, t_3, t_4) and d_2 consists of a bag of terms (t_2, t_5, t_6, t_7). Further, the user interest is specified as keyword t_2. For the sake of simplicity, we ignore smoothing and assume Q_m is only dependent on event language model of terms, i.e M_{et_e}. Hence, for a user interest t_2, the $Q_m(u,e)$ according to the EHD Model is equal to $\frac{2}{8}$. Whereas the individual contribution of a document d_1 and d_2 to Q_m is also equal to $\frac{1}{4}$ and their average contribution is $\frac{2}{4}$ (similar to EHD). But for a different structure of documents, the EHD model provides weaker estimates. (Refer figures 3.2 and 3.3).



Different contribution to the quality

FIGURE 3.2: Example: EHD Model makes a difference in case of different representations of document for an event



Same contribution to the quality

FIGURE 3.3: Example: EHD Model does not make a difference for a homogeneous cluster of documents

Hence, the above approach has suggested that by grouping documents into events, differences between representations of individual documents are, in effect, smoothed out. Therefore to incorporate the individual documents effect on the the quality we proposed another language model in the next section.

3.1.2 Language Model Based on ESD

In this approach events are ranked on the basis of a quality measure Q_m which is defined for a user interest u and an event e as

$$Q_m(u, e) = \frac{\sum_{\forall d \epsilon e} \log P(u|d)}{S_e}, \tag{3.8}$$

where S_e represents the size of the event, i.e., the number of documents in event e. The following table shows the notation that we follow in designing our ESD language models.

The basic idea of the **ESD** is to develop a language model for each document d in an event e to estimate the likelihood of a document being generated by user interests. The

d	document
e	event
T_d	Bag of tags for a document
T_{e_d}	Bag of terms for a document
λ, β	Smoothing parameters
$\{w_1, w_2, ..., w_m\}$	words in a user interest

TABLE 3.2: Notation - ESD Model

likelihood is defined as the probability that user interest u retrieves document d, i.e. $P(u|d)$. Then we subsitute the value of $P(u|d)$ in the equation (3.8) to calculate the quality measure for an event. In the following, we discuss the computation of $P(u|d)$ using language models.

Assuming a unigram language model for a document d, which states that keywords (w_1, w_2, w_3) in a user interest are independent of each other. With such a model, the problem is reduced to estimate $P(w_i|d)$.

$$P(u|d) = P(w_1|d)P(w_2|d)...P(w_m|d) = \prod_{i=1}^{m} P(w_i|d),$$

$$or,\ \log P(u|d) = \sum_{i=1}^{m} \log P(w_i|d). \tag{3.9}$$

Thus, according to the maximum likelihood estimator, $P(w_i|d)$ is given as,

$$P(w_i|d) = \frac{|\{w_i|w_i \in d\}|}{|d|},$$

One problem with this estimation is that an unseen word in a document d would get zero probability. More importantly, if a document is very small then the above described maximum likelihood estimation is generally not accurate. So in order to tackle these problems, we have to smooth the maximum likelihood estimation such that we do not assign zero probability to unseen words and improve the accuracy of the estimated language model in general. Therefore we incorporate the **Jelinek-Mercer smoothing model** [12], which is a combination of document language model M_d and collection language model M_c to each document d in an event e. Hence we obtain $P(w_i|d)$ as

$$P(w_i|d) = (1 - \lambda)M_d(w_i) + (\lambda)M_c(w_i) \tag{3.10}$$

$\lambda \in (0, 1)$ that controls the relative weight given to M_d and M_c. In our problem setting, a document is not simply a collection of terms but it is also comprised of rich bag of tags. Conceptually, the maximum likelihood of a user interest being generated as a **tag** is entirely different from the maximum likelihood of a user interest being generated as

a **term**. One can get better maximum likelihood estimations if we integrate the term and tag estimations differently into our language models. Hence, to incorporate tag and term importance or weights in a document language model $M_d(w_i)$ and in collection language model $M_c(w_i)$, we define them as

- $M_d(w_i)$ is a combination of $M_{dt}(w_i)$ and $M_{dt_e}(w_i)$, which are the maximum likelihood language models based on the count of w_i appearing in a document as a tag and a term respectively.

$$M_{dt}(w_i) = P(w_i|T_d) = \frac{|\{w_i|w_i \in T_d\}|}{|T_d|},$$

where T_d is a bag of tags for a document d.

$$M_{dt_e}(w_i) = P(w_i|T_{e_d}) = \frac{|\{w_i|w_i \in Te_d\}|}{|Te_d|}.$$

where T_{e_d} is bag of terms for a document d.

- Similarly, $M_c(w_i)$ is a combination of $M_{ct}(w_i)$ and $M_{ct_e}(w_i)$, which are maximum likelihood language models based on the count of w_i appearing in the collection as a tag and a term respectively.

$$M_{ct}(w_i) = P(w_i|T) = \frac{|\{w_i|w_i \in T\}|}{|T|},$$

where T is bag of tags obtained from collection of events.

$$M_{ct_e}(w_i) = P(w_i|T_e) = \frac{|\{w_i|w_i \in T_e\}|}{|T_e|}.$$

where T_e is bag of terms obtained from collection of events. We obtain $M_d(w_i)$ and $M_c(w_i)$ as

$$M_d(w_i) = (1-\beta)M_{dt}(w_i) + \beta M_{dt_e}(w_i), \; and \tag{3.11}$$

$$M_c(w_i) = (1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i). \tag{3.12}$$

where $\beta \in [0, 1]$ controls the relative weight given to the tag and term models. If β is 0 then the contribution of maximum likelihood estimation M_{dt_e} and M_{ct_e} is zero.

From equations (3.11) and (3.12), we obtain the probability of a word w_i in a user interest u of being generated by an event e as

$$\begin{aligned} P(w_i|d) = ((1-\lambda)((1-\beta)M_{dt}(w_i) + \beta M_{dt_e}(w_i))) \\ +(\lambda)((1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i))). \end{aligned} \tag{3.13}$$

Finally from equations (3.8) and (3.13), we obtain quality measure Q_m

$$\begin{aligned} \mathbf{Q_m(u,e)} = \sum_{d \epsilon e} \sum_{i=1}^{m} \{\log((1-\lambda)((1-\beta)M_{dt}(w_i) + \beta \\ M_{dt_e}(w_i)) + (\lambda)((1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i))\} \Big/ S_e. \end{aligned} \tag{3.14}$$

In order to derive the above quality measure, we have applied multiple smoothing methods to get better and effective retrieval results. Firstly, we smoothed the document model with the collection model using Jelinek Mercer Smoothing model [12], and then each document model and its respective collection model is smoothed using the tag model and term model. Moreover to improve the retrieval results, we integrated the S_e (the count of the number of documents in an event) component into the above quality measure. Specially, the integration of S_e will be useful on collections of events of hetrogenous sizes. For example, suppose we have one event in a collection containing 100 documents (out of which, 40 are only relevant documents) and the other event contains 50 relevant documents. Ignoring the division by S_e in this case may result that the former event quality measure is better than the later event which contains all relevant documentsevent. This is clearly undesirable. Therefore we divide by S_e to eliminate cases where some events can get unnecessary weightage due their size.

3.1.3 TFIDF Based on EHD

Despite of the popularity and effectiveness of the use of language models in the retrieval related tasks, there have been no concrete conclusion that *Language models outperforms traditional TFIDF models in every retrieval related task.* Therefore to validate our proposed retrieval approaches based on language models, we propose two other retrieval approaches using baseline as TFIDF. First retrieval method is designed using EHD model and the second method using ESD model. In this section, we discuss the TFIDF based on EHD model.

In this approach the quality measure for an event e and user interest u is defined as

$$Q_m(e,u) = \sum_{i=1}^{m} (1-\beta) Tg_{tfidf}(w_i,u) + (\beta) Tm_{tfidf}(w_i,u) \tag{3.15}$$

where β is a relative weight between $[0, 1]$ that controls TFIDF measure of w_i as tag (Tg_{tfidf}) and term (Tm_{tfidf}) for an event e respectively. Informally, the intuition behind the parameter β can be explained as follows; we know smoothing plays a critical role in the language models, but it can analogously be realized by term weighting in the TF*IDF model too. Therefore, we incorporated a relative weight β to control the tag and term importance in an event.

Thereafter, the TFIDF measure of a word w_i as a tag in an event e is given as

$$Tg_{tfidf}(w_i, q) = tgf_{w_i,e} \ * \ \log(\frac{n}{1 + tgdf_{w_i,E}}),$$

where n is the total number of documents in the collection. $tgdf_{w_i,E}$ is the number of documents in the collection that contains w_i as a tag. $tgf_{w_i,e}$ is the number of occurrences of w_i as a tag in an event e divided by total number of tags in an event e.

Also, the $TFIDF$ measure of a word w_i as a term in an event e is given as

$$Tm_{tfidf}(w_i, q) = tmf_{w_i,e} \ * \ \log(\frac{n}{1 + tmdf_{w_i,E}}),$$

where n is the total number of documents in the collection. $tmdf_{w_i,E}$ is the number of documents in the collection that contains w_i as a term. $tmf_{w_i,e}$ is the number of occurrences of w_i as a term in an event e divided by total number of terms in an event e.

Comparing it to the traditional $TFIDF$ model from the literature [14], the term frequency TF part is given by tags $tgf_{w_i,e}$ and terms $tmf_{w_i,e}$. The inverse document frequency IDF is measured both for tags ($\log(\frac{n}{1+tgdf_{w_i,E}})$) and terms ($\log(\frac{n}{1+tmdf_{w_i,E}})$). Here, the IDF part measure the general importance of the tag and a term which is obtained by dividing the total number of documents by the number of documents containing the tag and term respectively, and then taking their logarithms.

3.1.4 TFIDF Based on ESD

Similarly in the $TFIDF$ retrieval approach based on ESD, we compute TF*IDF tag and term measure for each document in an event e and then sum up the TFIDF values for member documents of an event e to obtain the the quality measure. It is given as

$$Q_m(e, q) = \sum_{d \in e} \sum_{i=1}^{m} (1 - \beta) Tgd_{tfidf}(w_i, q) + (\beta) Ted_{tfidf}(w_i, q) \Big/ S_e \tag{3.16}$$

where β is a relative weight between $[0, 1]$ that controls TFIDF measure of w_i as tag (Tgd_{tfidf}) and term (Ted_{tfidf}) for a document d. S_e is the size of an event in terms of number of documents in e.

The TFIDF measure of word w_i as tag in document d is given as

$$Tgd_{tfidf}(w_i, q) = tgf_{w_i,d} * \log(\frac{n}{1 + tgdf_{w_i,E}}),$$

The TF*IDF measure of word w_i as term in document d is given as

$$Ted_{tfidf}(w_i, q) = tef_{w_i,d} * \log(\frac{n}{1 + tedf_{w_i,E}}),$$

n is the total number of documents in the collection. $tgdf_{w_i,E}$ and $tedf_{w_i,E}$ is the number of documents containing w_i as a tag and a term in collection of events E. $tef_{w_i,d}$ is the number of occurrences of w_i as a term in a d divided by total number of terms in a d. Also, the $tgf_{w_i,d}$ is the number of occurrences of w_i as a tag in a d divided by the total number of tags in a d.

3.2 Comparing Retrieval Approaches

1. The language model on **ESD** is viewed as a mixture of the following three sources, which are tags in a document, terms in a document, and the collection of documents from the entire collection of events. And the language model on EHD is comprised of sources; tags in event, terms in event, and the collection of events. The **EHD** model, while being simple and intuitive, may have a number of problems such as the event model illustrated as cases, below.

   Case 1 : Suppose there are two events e_1 and e_2 and each event is a cluster of documents as shown in Figure 3.4. Further, each document contains a bag of terms, designed as round dots in the Figure 3.4. Now the question is that for a user interest (specified as colored dot in the Figure 3.4), which event will the user prefer? Let's rely on retrieval models based on ESD and EHD to retrieve the best event for a specified user interest. The retrieval methods based on EHD model will equally score both events because their model in independent of distribution of the terms in the user interest inside individual documents of an event. But the retrieval model based on ESD will favor event e_1 over e_2 because the model captures the distribution of keywords in user interest for each document.

   Case 2 : Suppose there are two events e_1 and e_2 and a user interest contains multiple keywords (shown as rectangle and round dot in the Figure 3.5). Both events contains documents that contain the terms matching to keywords in a user

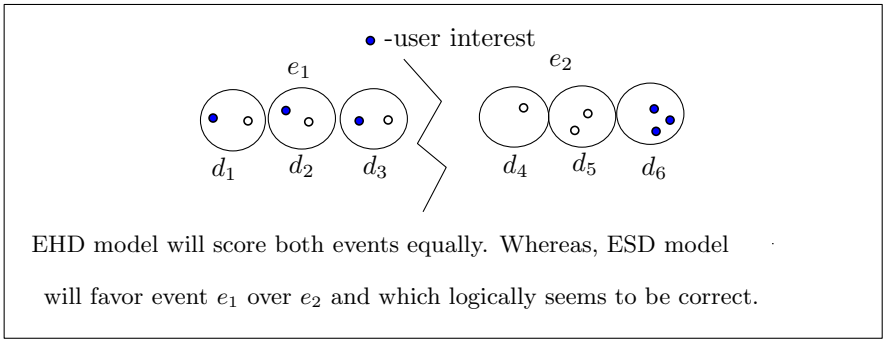


FIGURE 3.4: Case 1 : Difference in results obtained by ESD and EHD

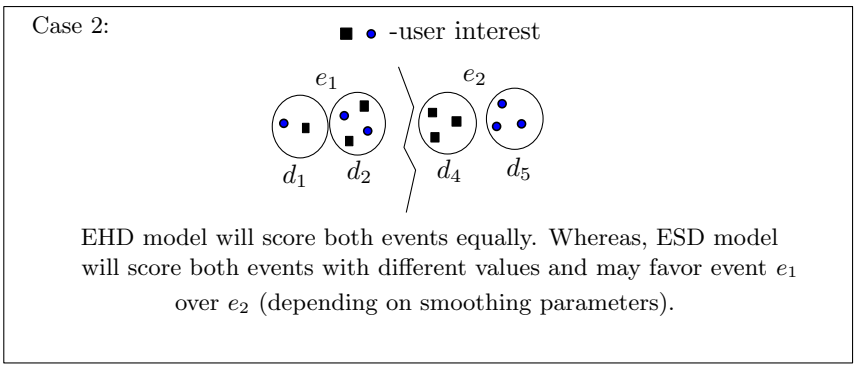


FIGURE 3.5: Case 2 : Difference in results obtained by ESD and EHD

interest. But documents (d_1 and d_2) of an event e_1 are equally dominated by the keywords. In event e_2, the document d_5 is dominated by one keyword (represented as rectangle) and the other document d_5 is dominated by the other (represented as round dot). In this case also, EHD will score both events equally and the ESD model will provide us with different scoring values. It is inferred from the above cases that the ESD model better captures the characterstics the documents that comprise the event than EHD model. Therefore, atleast theoretically, retrieval methods based on the ESD model looks more promising.

2. Language model approaches are more effective than traditional TFIDF retrieval approaches according to the research work by Ponte and Croft [9]. Also in the book [12], some interesting relation is derived between smoothing in language models and TFIDF weighting which explains that the estimates provided by language models are more accurate than TFIDF.

3. Actually, it is very hard to comment that which retrieval model is the best. The results obtained by the retrieval models are very much dependent on the input data source, that is why we report experimental studies in Chapter 7.

3.3 Summary

In this chapter we proposed various retrieval approaches to solve the problem of personalization of emerging topics on the basis of user interest. In Chapter 8, we evaluate all these retrieval approaches for different user interests. In the next chapter, we discuss some techniques to expand or augment user interests in order to refine the personalized results.

Chapter 4

Refining Personalized Events using Local Context Analysis

We aim to augmenting user interests in order to re-rank the personalized events. Our approach is adapted from Local Context Analysis [4], [3]. It is a query reformulation technique, which selects expansion terms based on co-occurrence with the query terms within the top ranked documents. The expansion terms selected by local context analysis are called concepts.

Local Context Analysis: Let $w_1, w_2, ..., w_m$ are the query terms of query Q to be expanded, N be the number of documents in collection C, N_c the number of documents that contain concept c, S be the set of top-ranked documents as $S = \{d_1, d_2, ..., d_n\}$, and $co(c, w_i)$ the number of co-occurrences between c and w_i in S. They used the following $codegree(c, w_i)$ metric to measure the degree of co-occurrence of c with w_i.

$$\begin{aligned} codegree(c, w_i) &= \log(co(c, w_i) + 1)\frac{idf(c)}{\log(n)}. \\ co(c, w_i) &= \sum_{d \in S} tf(c, d)tf(w_i, d), \\ idf(c) &= min(1.0, log(N/N_c)/5.0). \end{aligned} \tag{4.1}$$

where $tf(c, d)$ and $tf(w_i, d)$ are the frequencies of c and query term w_i in document d, respectively. The metric takes into account the frequency of c in the whole collection $idf(c)$ and the number of co-occurrences between c and w_i in the top ranked set $co(c, w_i)$. Our proposed method for user interest expansion is based on this metric.

We should explain why we adapted the above co-occurrence metric for our proposed method and not other well-known metrics such as EMIM (expected mutual information measure) [28], cosine [14], and so forth. One reason we choose not to use these metrics

is that they are designed to measure corpus wide co-occurrence, and it is not clear how to adapt them to measure co-occurrence in the top ranked documents. The other reason is that we want to explicitly bias against high-frequency concepts, but available metrics cannot do that.

In our proposed method, the codegree metric is generated using language models but not the TF*IDF measure as used in classical LCA. Moreover, we expand user interest by not only appending terms but tags, too. These are the main differences of our proposed approach from LCA. In the next section, we discuss in detail our method to re-rank personalized results by reformulating user interests.

4.1 Re-ranking Personalized Results

Our proposed method follows the following simple steps for re-ranking personalized results.

1. Extract *popular terms and tags* (Section 4.1.1) from the events related to personalized results.
2. Choose m terms and n tags on the basis of popularity to compute the *co-occurrence of popular terms and tags* (Section 4.1.2) with the words in the initial user interest. The values of m and n are chosen experimentally.
3. Re-evaluate the *augmented user interest* to generate the new (refined) (Section 4.1.3) ranking of the personalized results.

4.1.1 Extracting Popular Terms or Tags

Popular terms or tags : A term is defined as popular if the term count is more than the threshold θ. Similarly a tag is defined as popular if the tag count is larger than threshold θ. The term count and tag count is the number of times a given term and tag appears in the documents linked to the events in the personalized results, respectively.

We compute the popularity of all terms and tags from term and tag usages of all the documents linked to the events related to the personalized results. Then we extract m terms and n tags which have popularity measure more than threshold θ. The m, n and θ values can be decided experimentally.

4.1.2 Computing Co-occurrence

If the retrieval method used for personalization is the EHD language model, refer Section 3.1.1 for details related to constructing a language model. The *codegree* metric for a interest u $(w_1, w_2....)$ with popular term or tag p and for a given event e is given as,

$$codegree(p, u, e) = \log P(p|e) + \sum_{i=1}^{m} \log P(w_i|e) \tag{4.2}$$

Let u be a user interest and e an event that belongs to the result set S. M_{et}, M_{et_e}, M_{ct} and M_{ct_e} are maximum likelihood language models based on the $w_i's$ or p occurrence as tag in an event e, as term in an event e, as tag in collection S and as term in collection S,respectively. The $logP(w_i|e)$ from Equation (4.2) is given as,

$$\begin{aligned} \log P(w_i|e) = \log(((1-\lambda)((1-\beta)M_{et}(w_i) + \beta M_{et_e}(w_i))) \\ +(\lambda)((1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i))). \end{aligned} \tag{4.3}$$

Similarly, the $\log P(p|e)$ can also be formulated.

If the retrieval method used for personalization is the ESD language model, refer Section 3.1.2 for details related to constructing a language model. The *codegree* metric for interest u $(w_1, w_2, ...)$ with popular term or tag p is given as, is given as,

$$codegree(p, u, e) = \frac{\sum_{d \in e}(\log P(p|d) + \sum_{i=1}^{m} \log P(w_i|d))}{S_e}. \tag{4.4}$$

where S_e (count of documents in an event e). For a user interest u and event e in the result set S; M_{dt}, M_{dt_e}, M_{ct} and M_{ct_e} are maximum likelihood language models. The M_{dt} and M_{dt_e} are the language models that estimates the maximum likelihood of $w_i's$ or p (4.5) as tag or a term in a document d. The M_{ct} and M_{ct_e} are the language models that estimate the maximum likelihood of $w_i's$ or p (4.5) as tag or a term in a collection. The $logP(w_i|d)$ from Equation (4.4) is given as

$$\begin{aligned} \log P(w_i|d) = \log((1-\lambda)((1-\beta)M_{dt}(w_i) + \beta M_{dt_e}(w_i)) \\ +((\lambda)((1-\beta)M_{ct}(w_i) + \beta M_{ct_e}(w_i)))). \end{aligned} \tag{4.5}$$

Without Expansion -Events retrieved by ESD-LM

MURDERS AND ATTEMPTED MURDERS**SHOOTINGS
ATTACKS ON POLICE**RELIGIOUS CULTS
ATTACKS ON POLICE**SHOOTINGS

With Expansion

SECURITY AND WARNING SYSTEMS**WORLD TRADE CENTER (NYC)
SECURITY AND WARNING SYSTEMS**EXPLOSIONS
ISLAM**EXPLOSIONS

FIGURE 4.1: Improving Personalized Results by Expansion

4.1.3 Refined Ranking

We compute the *codegree* metric for every event in the result S for a given set of popular terms and tags. We then sum up the values obtained by *codegree* metric for m popular terms and n popular tags (say it comprises a set P') in order to get the refined score of an event.

$$RefinedScore(e) = \sum_{p \in P'} codegree(p, u, e).$$

4.2 Discussion

We know that personalization models can be best evaluated by a user study, however due to the time constraints, we do not have any evaluation from the users for validating the work related to user interest expansion technique. Therefore, to provide the better intuition of our work, we would like to discuss our observation on how the personalized events retrieved by the $ESD-LM$ are affected by the user interest expansion technique.

We specified the user interest as *Terrorist Attack* to retrieve the relevant event from the collection of events detected during the timeline : Year-1993, Month-Feb, and Week-4. Then with the aim to improve the results, we applied the user interest expansion technique. The initial user interest *"Terrorist Attack"* is expanded to *"Terrorist Attack BOMBS WORLD TRADE CENTER EXPLOSIONS"* and our observation suggests that the expansion technique helps in improving the personalized results. It can be observed from Figure 4.1 only that the Top-3 events returned after expansion are better than the personalized results without expansion.

More such instances are found for different user interests, like, *Airline Delays*, *Germany Politics* and many other, which make us believe that the proposed expansion technique really helps in improving the personalized retrieved results.

Chapter 5

Diversity in Personalized Events

In cases when there are numerous relevant events which are highly redundant with each other, i.e., containing duplicate information, we must think of an approach for eliminating redundancy from the result set. Suppose a user interest is "Stock Market", our retrieval model returns ranked list of the top-10 events (more if requested) and then the user examines the top ranked event. It is *Mutual Funds**Stocks and Bond* and contains documents related to *investment in mutual funds and in stocks* which means it is a very relevant event. The next event is *Investment Strategies**Mutual Funds* and contains documents related to *investment in mutual funds.* Though the event is relevant, it contains documents which were already seen in the previous event. Hence the later event does not provide any new information to a user. Most events in the top-k set repeat information already contained in the previous ones, and the user get dissatisfied with the results. In our work we present an approach to *diversify* results in order to minimize the risk of dissatisfaction of the average user.

5.1 Related Work

By avoiding duplicated and nearby duplicated documents, diversity featured search systems try to meet user information needs by generating more diverse results. The diversity problem is addressed in literature as **Maximum Diversity Problem** [29] and **MaxMin Diversity Problem (MMDP)** [30], [31]. In information retrieval community-Carbonell and Goldstein proposed **Maximal Marginal Relevance** [32] algorithm to realize diversity by selecting documents that are both relevant and least similar to the documents already chosen.

Maximum Diversity Problem, the problem was introduced by Glover [29], and is strongly NP-hard [33], [34], [31]. This problem has a significant number of practical

applications such as environmental balance, telecommunication services, and genetic engineering. For a given set N of n objects and a matrix d_{ij} providing the diversity between each pair of objects i and j belonging to N, the Maximum Diversity Problem (MDP) consists in finding a set $M \subset N$ of m objects such that the sum of their pairwise distances is maximum. In this problem setting, it is assumed that the diversity matrix is symmetric (i.e. $d_{ij} = d_{ji}$ for all $i, j \in N$). The other assumption is that $d_{ii} = 0$ for all $i \in N$ and $x_i = 1$ if element $i \in N$ belongs to the solution M, $x_i = 0$ otherwise.

$$\begin{aligned} &Objective: \\ &\max \frac{1}{2} \sum_{i \in N} \sum_{j \in N} d_{ij} x_i x_j \\ &Constraints: \\ &\sum_{i \in N} x_i = m, \\ &x_i \in \{0,1\}, i = 1, 2, ..., n. \end{aligned} \tag{5.1}$$

In literature, the diversity problem is also studied as the MaxMin Diversity Problem (MMDP). The problem is NP-hard shown by Erkut [30] and Ghosh [31]. It consists of selecting a subset M of m elements from a set N of n elements in such a way that the minimum distance between the chosen elements is maximized. It assumes that the diversity matrix is symmetric (i.e. $d_{ij} = d_{ji}$ for all $i, j \in N$), $d_{ii} = 0$ for all $i \in N$ and $x_i = 1$ if element $i \in N$ belongs to the solution M, $x_i = 0$ otherwise.

$$\begin{aligned} &Objective: \\ &\max \min_{i,j \in N} d_{ij} x_i x_j \\ &Constraints: \\ &\sum_{i \in N} x_i = m, \\ &x_i \in \{0,1\}, \ i = 1, 2, ..., n. \end{aligned} \tag{5.2}$$

Apart from some mathematical programming approaches and from greedy and stingy heuristics [29][35] (heuristics based on branch and bound method or simplex methods), the literature on the MDP and $MMDP$ consists of several *Greedy Randomized Adaptive Search Procedures (GRASP)* [36], [37].

Carbonell and Goldstein motivate **Maximal Marginal Relevance (MMR)** [32] with the intention to include diversity into the ranking of documents to prevent presentation of redundant information. Let N be the ranked list of documents, S the set of documents

in N which are already retrieved, N/S is the set difference, i.e, the set of yet unselected documents in S. The Sim_1 is the similarity metric used in relevance ranking between documents and a query and Sim_2 is the similarity between the pair of documents. Moreover Sim_2 can be the same metric as Sim_1 or a different one.

$$MMR = \arg \max_{D_i \in N/S} [\lambda(Sim_1(D_i, Q)) - (1 - \lambda)(\max_{D_j \in S} Sim_2(D_i, D_j))] \tag{5.3}$$

Given the above definition, MMR computes incrementally the relevance ranked list when the parameter $\lambda = 1$, and computes a maximal diversity ranking among the documents in R when $\lambda = 0$. A linear combination of both of these criteria is optimized for the intermediate values of λ in the interval $[0, 1]$.

While most of the research work on diversity in IR is concerned with ranking of documents, our work addresses the problem of diversity in emerging topics. Though the MMR algorithm can be easily extended to our work of diversify personalized emerging topics; it has some drawbacks, like, in MMR there is a constraint of tuning the parameter λ which can be obtained through experiments on different data sets.

The formulations of diversity problems as MDP and $MMDP$ were studied and evaluated on a comprehensive set of benchmark instances representative, such as, *SOM* which consists of 70 matrices with random numbers between 0 and 9 generated from an integer uniform distribution, *GKD* which consists of 145 matrices for which the values were calculated as the Euclidean distances from randomly generated points with coordinates in the 0 to 10 range and *MDG* which consists of 100 matrices with real numbers randomly selected between 0 and 10 from a uniform distribution. A brief description of the origin and the characteristics of these set of instances and more is given in [38], [39].

According to our knowledge these methods are not studied or reformulated for computing maximal diversity ranking for events. In our work we aim to do so by redefining the problem of diversity for our problem setting, i.e., to extract the topics obtained by one of our retrieval models (refer Chapter 4) which are novel and distinct, and also highly matches the interest of user. We define the problem as **Maximum Diversity Quality Problem** and **Maximum Min-Diversity Quality Problem** because we are concerned in retrieving events which are highly relevant (maximum quality) and distinct from other events in the ranked list (maximum diversity). Moreover, in our work we are not restricted to Euclidean distance as the only diversity measure. A detailed classification of diversity measures is discussed in Chapter 7. The main motivation of the following work is that increasing the dissimilarity among the retrieved results increases the satisfaction in a user for the presented results.

5.2 Maximum Diversity Quality Problem

Given a set N of n events for which diversity measure d_{ij} is defined for every pair of events (i, j) such that $d_{i,j} \in [0, 1]$. In our problem setting, the $d_{i,j}$ is symmetric (i.e., $d_{ij} = d_{ji}\ \forall i, j \in N$), $d_{ii} = 0$ for all $i \in N$, and quality measure $q_i \in (0, 1)$ is defined for every event i.

The **Maximum Diversity Quality Problem (MDQP)** consists in determining a subset $M \subset N$ of cardinality m, such that the sum of the product of diversity and quality measure of events in M is maximum. The MDQP can then be formulated as the following *quadratic zero-one integer program*, where $x_i = 1$ if i in N belongs to the solution M, otherwise $x_i = 0$

$$\begin{aligned} &Objective: \\ &\max\ z = \sum_{i \in N} \sum_{j \in N} (d_{ij} x_i x_j) * q_i. \end{aligned} \tag{5.4}$$

$$\begin{aligned} &Constraints: \\ &\sum_{i \in N} x_i = m, \\ &x_i \in \{0, 1\}, \forall i \in N. \end{aligned}$$

5.2.1 Linear Integer Program

Glover and Wolsey [35] show how to convert a 0-1 polynomial programming problem into a 0-1 linear programming problem by replacing the product of some variables with a new variable satisfying some additional constraints. Kuo et al. [29] used this transformation to convert quadratic formulation of Maximum Diversity Problem to linear program. We also apply same transformation and replace the product $x_i x_j$ in Equation (5.4) with a new variable y_{ij} and obtain the **linear integer program** below in which the additional constraints have been included.

$$\begin{aligned} &Objective: \\ &\max\ z = \sum_{i \in N} \sum_{j \in N} (d_{ij} y_{ij}) * q_i). \end{aligned}$$

$$Constraints:$$

$$\sum_{i \in N} x_i = m, \tag{a}$$

$$y_{ij} \leq x_i,\ \forall i, j \in N, \tag{b}$$

$$y_{ij} \leq x_j, \ \forall i, j \in N, \tag{c}$$

$$y_{ij} \geq x_i + x_j - 1, \ \forall i, j \in N, \tag{d}$$

$$x_i \in \{0, 1\}, \ \forall i \in N, \tag{e}$$

$$y_{ij} \in \{0, 1\}, \ \forall i, j \in N. \tag{f}$$

The constraints (b), (c) and (d) guarantee that the auxiliary variable y_{ij} equals 1 if both objects i and $j \in M$, 0 otherwise. The above linear integer program can be solved by some standard LP Solvers such as Cplex [23], GNU Linear Programming Kit (GLPK). The major constraint in using the solver for the diversity problem is time. The solution from the solver will take an excessively long time. Therefore the use of CPLEX solver (though it is considered to be one of fastest and efficient solver) for retrieving m diverse events or topics from a set N of n events is not at all recommended. We propose the greedy heuristic as an alternative efficient solution to the Maximum Diversity and Quality Problem.

5.2.2 Greedy Heuristic for MDQP

In the following, we explain a greedy heuristic to solve $MDQP$ problem. Conceptually, in the greedy approach we make a locally optimal choice at each step in the hope of finding the global optimal solution.

For $MDQP$ problem, the contribution of each event $i \in N$ to solution set M is defined as

$$D_i = \sum_{j \in M} d_{ij} * q_i.$$

The initial solution $M^{(0)} \subset N$ is composed of an event i such that it maximizes the product of diversity measure (d_{ij}) and quality measures q_i such that

$$M^{(0)} = \arg \max_{i,j \in N} (d_{ij} * q_i). \tag{5.5}$$

A feasible solution containing m elements is built from $M^{(0)}$ by adding one element at a time. At h^{th} step, the element $k^{(h)}$ to be added is chosen as

$$m^{(h)} = \arg \max_{i \in N \setminus M^{(h-1)}} D_i.$$

It results in the following solution,

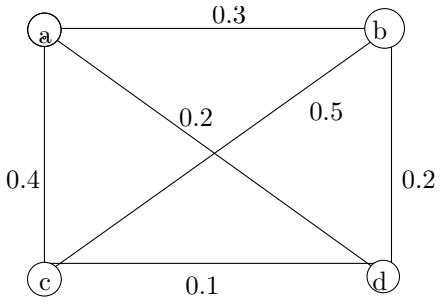
$$M^{(h)} = M^{(h-1)} \cup m^{(h)},$$

and our objective function z at iteration h is equal to :

$$z^{(h)} = z^{(h-1)} + D_{m^{(h)}}.$$

At the end of each iteration, we have to update D_i for each $i \in N \backslash M^{(h)}$ and it is done by adding $(d_{i\,k^{(h)}} * q_i)$ to D_i. Each iteration requires $O(n)$ time, so that the overall procedure takes $O(mn)$ time.

For illustration, consider a set of 4 events (a, b, c, d) represented as nodes as shown in Figure 5.1 where edge specify the diversity measure $(d_{i,j})$ between a pair of events (i, j). We made an assumption that the quality of each event (q_i) is 1.0. So at the first step, we choose an event which has maximum $(d_{ij} * q_i)$, which is either b or c. Assume, we choose b and at the second step, we choose an event which has maximum $(d_{ib} * q_i)$, it is c. Then at the last step, we choose event a as it provides the better contribution to the solution set in comparison to event d.



b-c	$0.5 * 1.0 = 0.5$	b
c-b	$0.5 * 1.0 = 0.5$	c
a-(b,c)	$(0.3) * 1.0 = 0.3$	a
d-(b,c)	$(0.1) * 1.0 = 0.1$	

a, b , c and d are events; edge specify the diversity for a pair of event and quality of each event is 1.0

FIGURE 5.1: Illustration of the Greedy Solution to the Maximum Diversity Quality Problem

5.2.3 Discussion

The non transitivity of diversity relation may provide less accurate solution obtained through the MDQP. By non transitivity of the diversity relation, we mean that if an event a is diverse from b by value x and an event b is diverse from c by value y, it does not mean that a is diverse from c by value $x + y$. This property may result in the weak solution in some cases, for illustration, suppose there are 4 events (a, b, c, d), out of which

2 are already in the solution set M, and the problem is to add one more event to a set M.

According to the solution strategy for MDQP, an event a is added but the diversity measures suggest that the addition of b would have been more beneficial as shown in Figure 5.2. Therefore, to address such cases and to provide a better solution, we propose the alternative solution to the diversity problem by defining the problem as *Maximum Min-Diversity Quality Problem*. It is explained in the following section.

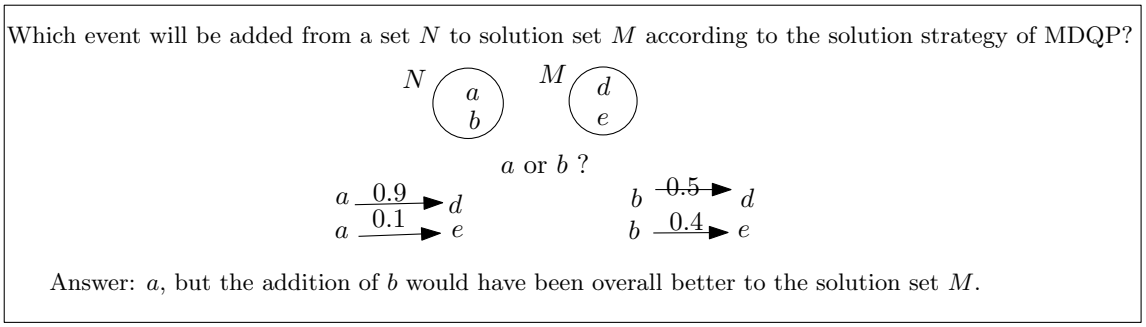


FIGURE 5.2: Illustration: Weak solution obtained by MDQP in some cases

5.3 Maximum Min-Diversity Quality Problem

The *Maximum Min-Diversity Quality Problem* consists of selecting a subset M of m events from a set N of n events in such a way that the sum of $(\min_{ij \in M} d_{ij}) * q_i$ for all chosen events is maximized. The d_{ij} is the diversity between events i and j and it is symmetric ($d_{ij} = d_{ji}$), q_i is the quality of an event i, x_i takes the value 1 if event e_i is selected and 0 otherwise. Hence the objective of the problem is realized as the following quadratic integer program,

$$\max\ z = \sum_{i \in N}((\min_{j \in N, j \neq i} d_{ij} x_i x_j) * q_i). \tag{5.6}$$

$$s.t. \sum_{i \in N} x_i = m,$$

$$x_i \in \{0, 1\}\ \forall i \in N.$$

5.3.1 Linear Integer Program

The linear integer formulation of the $MMDP$ quadratic objective function is described below

Objective :

$$\max\ z = \sum_{i \in N} (w_i y_{ij}) * q_i.$$

Constraints :

$$0 < w_i \leq d_{ij} + K * (1 - y_{ij}),\ \forall j \neq i, \tag{g}$$

$$\sum_{i \in N} x_i = m, \tag{h}$$

$$y_{ij} \leq x_i,\ \forall i, j \in N, \tag{i}$$

$$y_{ij} \leq x_j,\ \forall i, j \in N, \tag{j}$$

$$y_{ij} \geq x_i + x_j - 1,\ \forall i, j \in N, \tag{k}$$

$$x_i \in \{0, 1\},\ \forall i, j \in N, \tag{l}$$

$$y_{ij} \in \{0, 1\},\ \forall i, j \in N. \tag{m}$$

where in constraint (g) K is a sufficiently large parameter chosen to capture the minimum value in w_i and condition $(w_i > 0)$ avoids to choose an event i with $d_{ij} = 0$. The constraints (j), (k) and (l) guarantee that the variable y_{ij} equals 1 iff both events i and $j \in M$ and 0 otherwise.

5.3.2 Greedy Heuristic

The contribution of each event i to the solution set M is given as

$$D_i = (\min_{j \in M} d_{ij}) * q_i. \tag{5.7}$$

In the first step $M^{(0)} \subset N$ is composed of an event that maximizes the product of diversity measure d_{ij} and quality measure q_i such that

$$M^{(0)} = \arg \max_{i,j \in N} d_{ij} * q_i. \tag{5.8}$$

A solution containing m elements is built from $M^{(0)}$ by adding one element k at a time. At the h^{th} step, the element $k^{(h)}$ to be added is chosen as

$$m^{(h)} = \arg \max_{i \in N \setminus M^{(h-1)}} D_i.$$

It results in the following solution

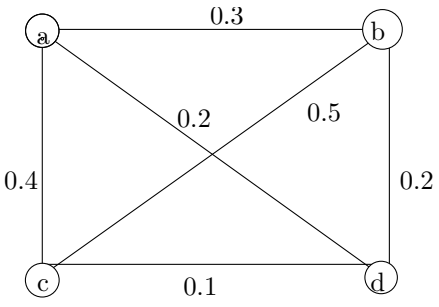
$$M^{(h)} = M^{(h-1)} \cup m^{(h)},$$

and our objective function z at iteration h is equal to

$$z^{(h)} = z^{(h-1)} + D_{m^{(h)}}.$$

And at each iteration we update D_i for each $i \in N \backslash M^{(h)}$ and it is done by updating D_i to $\min_{\forall j \in S}(d_{i\ k^{(h)}})$, where $S \subset M$ at step h. Each iteration requires $O(n)$ time, so that the overall procedure is $O(mn)$.

For illustration, consider a set of four events (a, b, c, d) represented as nodes as shown in Figure 5.3 where edge specify the diversity measure $d_{i,j}$ between a pair of events (i, j). We made an assumption that the quality of each event q_i is 1.0. So at the first step, we choose an event which has maximum $d_{ij} * q_i$, which is either b or c. Assume, we choose b and in the second step, we choose an event which has maximum $d_{ib} * q_i$, which is c. Then in the last step, we choose an event that provides better contribution, i.e., $(\min_{j \in M} d_{ij}) * q_i$ to the solution set.



b-c	$0.5 * 1.0 = 0.5$	b
c-b	$0.5 * 1.0 = 0.5$	c
a-(b,c)	$(0.3) * 1.0 = 0.3$	a
d-(b,c)	$(0.1) * 1.0 = 0.1$	

a, b , c and d are events; edge specify the diversity for a pair of event and quality of each event is 1.0

FIGURE 5.3: Illustration of the Greedy Solution to the Maximum Min-Diversity Quality Problem

5.3.3 Discussion

The use of min approximation guarantees that the event added in each iteration (Greedy Heuristic) is atleast diverse from the events in the solution set, by some value, say x. In other words, it provides the upper bound of the diversity measure of an event to the remaining events in the solution set. As the diversity problem in NP-hard, it is hard to say which approximation (Max sum in case of MDQP and Max sum-Min in case of

MMDQP) will work better. Actually, the nature of the dataset in terms of the diversity measures can help in deciding which approximation will work better. We postpone the discussion on which approximation is better as part of future experimental study.

5.4 Alternatives

The one possible alternative in formulating the objective function of the **Maximum Diversity Quality Problem** can be obtained by changing the product operator to sum as,

$$\begin{aligned} &Objective: \\ &\max\ z = \sum_{i \in N} \sum_{j \in N} (d_{ij} x_i x_j + q_i x_i). \\ &Constraints: \\ &\sum_{i \in N} x_i = m, \\ &x_i \in \{0,1\},\ \forall i \in N. \end{aligned} \tag{5.9}$$

where d_{ij} is a diversity measure for a pair of events, q_i is a quality measure of an event and $x_i \in \{0,1\}$. The greedy solution for this objective function is built by adding an event i which has maximum value of $d_{ij} + q_i$ in the first iteration. Then in each of the next iterations choose an event which provides the maximum value of $D_i = \sum_{j \in S} d_{ij} + q_i$ where $S \subset M$ and append the event i to the solution set M. Moreover, at the end of each iteration h we update D_i to $d_{i\,k^{(h)}}$ where $k^{(h)}$ is the event added during the h^{th} iteration.

Similarly, the **Maximum Min-Diversity Quality Problem** objective function can also be formulated by replacing the product to sum as we did in the $MDQP$.

$$\begin{aligned} &Objective: \\ &\max\ z = \sum_{i \in N} ((\min_{j \in N}\ d_{ij} x_i x_j) + q_i x_i). \\ &Constraints: \\ &\sum_{i \in N} x_i = m, \\ &x_i \in \{0,1\}; \forall i \in N. \end{aligned} \tag{5.10}$$

The greedy solution for this objective function is built by adding an event i which has maximum value of $d_{ij} + q_i$ in the first iteration. The contribution of each event i to the

solution set M is given as $D_i = \min_{j \in M} d_{ij} + q_i$. Then in each next iteration append the event i to the solution set M which has maximum value of D_i . Moreover at the end of every iteration h we update the D_i to $\min_{\forall j \in S}(d_{i\ k^{(h)}}$ where $S \subset M$ in step h and $k^{(h)}$ is the event added during the h^{th} iteration.

In this chapter, we discussed the solutions to eliminate redundant events from the result set such that there is a variety in the results presented to the users. In the next Chapter 6, we discuss the possible diversity measures related to our work.

Chapter 6

Diversity Measures

6.1 Classification of Diversity Measures

The diversity measures are classified as the following.

1. *Diversity - Documents*: Let D_i, D_j be a set of documents representing events e_i and e_j then the diversity measure d_{ij} for these pair of events is given as,

$$d_{ij} = 1 - \frac{|D_i \cap D_j|}{|D_i \cup D_j|},$$

   where $\frac{|D_i \cap D_j|}{|D_i \cup D_j|}$ is a similarity measure for a pair of events in terms of documents. It is also known as *Jaccard similarity coefficient* [40].

2. *Diversity - Taxonomic Categories*: Let e_t, $e'_{t'}$ are sets of taxonomies representing events e_i and e_j then the diversity measure d_{ij} for these pair of events is given as,

$$d_{ij} = 1 - SimEvents(e_t, e'_{t'}).$$

   where $SimEvents(.,.)$ provides the similarity measure for a pair of taxonomies.

The *Diversity-Documents* diversity measures is straightforward to compute, whereas, the *Diversity-Taxonomic Categories* diversity measure has some additional constaints due to the structure of taxonomic categories. Therefore, in the following, we discuss the computation of this measure.

6.2 Diversity - Taxonomic Categories

The diversity for a pair of events e and e' in terms of taxonomic categories is given as $1 - SimEvents(e_t, e'_{t'})$ where $SimEvents(e_t, e'_{t'})$ is a similarity metric, e_t and $e'_{t'}$ is a set of taxonomic categories defined for events e and e' respectively.

$$d_{ij} = 1 - SimEvents(e_t, e'_{t'}), \tag{6.1}$$

where i refers event e, j refers event e'. Though there are some standard metrics to compare the similarity of sets, like; Jaccard index, Dice's coefficient (introduced by Dice in 1945 [40]), etc. But these methods can not be directly applied to compute the similarity of sets of taxonomic categories due to the structure of taxonomic category. The taxonomic category is arranged in a hierarchical structure. At the top of this structure is a single classification, the root node, that applies to all concepts defined under that root node. Examples of taxonomic category obtained from New York Times Archive: **Top/Features/Travel/Guides** where *Top is the root node and Features, Travel and Guides are the concepts classified under the root node.*

Given a pair of sets of taxonomic categories $e_t = \{t_0, t_1, ...\}$ and $e'_{t'} = \{t'_0, t'_1, ...\}$ defined for events e_t and $e'_{t'}$ respectively; each taxonomic category consists of a set of concepts, e.g., $t_i = \{c[0], c[1], ...\}$, $t'_j = \{c'[0], c'[1], ...\}$. Now, either t_i and t'_j are totally similar (the hierarchy (sequence) of concepts of t_i is similar to t'_j), partially similar (the hierarchy of concepts of t_i partially overlaps t'_j or vice versa) or totally dissimilar (the hierarchy of concepts of t_i is entirely different from t'_j (excluding the root node from the hierarchy)). For example, Figure 6.1 elaborates all three cases.

similar pair, partial similar dissimilar pair of taxonomy categories

Top/Features/Travel/Guides/Destinations	Top/Features/Travel/Guides/Destinations
Top/Features/Travel	Top/Features/Travel/Guides/Destinations
Top/News	Top/Features/Travel/Guides/Destinations

TABLE 6.1: Example of taxonomic categories

6.3 Estimation of SimEvents Measure

The Algorithm 1 $TaxSim$ is used to evaluate the semantic similarity between t_i and t'_j. Intuitively, one key to the similarity of two categories is the extent to which they share information in common and which is determined by the common sequence of concepts. The edge counting method is referred to determine the similarity of two categories. The Algorithm $TaxSim$ returns 1 if t_i and t'_j are similar or their partial similarity is more than θ, otherwise it returns 0.

Algorithm 1 TaxSim: Calculate similarity of two taxonomic categories (t_i, t'_j)

Input: $t_i = \{c[0]..., c[n]\}, t'_j = \{c'[0], ..., c'[m]\}$, θ
Output: {0,1}/*1-similarity is more than θ, otherwise 0*/

```
s ⇐ 0
k ⇐ 0
while (k ≤ n) ∪ (k ≤ m) do
  if c[k] = c'[k] then
    i ⇐ k + 1
  else
    s = (k−1)/(n+m−k+1)
    if s ≥ θ then
      s = 1
    else
      s = 0
    end if
  end if
end while
return s
```

Algorithm 2 SimEvents : Count of similar taxonomies between events $(e_t, e'_{t'})$

Input: e_t, $e'_{t'}$
Output: ST

```
for all t ∈ e_t do
  for all t' ∈ e'_t do
    ST=ST+TaxSim(t, t')
  end for
end for
return ST/(E+E') /*E and E' is the total number of taxonomic categories in e_t and e'_t' respectively*/
```

Set of taxonomic categories of an event e : e_t

t_0	Top/Classifieds/ Job Market/ Job Categories/Banking, Finance and Insurance
t_1	Top/Features/Travel/Guides/Destinations/Asia

Set of taxonomic categories of an event e' : $e'_{t'}$

t'_0	Top/Features/Travel/Guides/Destinations
t'_1	Top/Features/Travel/Guides/Destinations/Central and South America

TABLE 6.2: Sets of taxonomic categories e_t and $e'_{t'}$

In order to calculate $SimEvents(e_t, e'_{t'})$ (see Algorithm 2), we call the procedure $TaxSim$ for every taxonomic category in e_t with every taxonomic category in $e'_{t'}$. It returns the similarity measure for a pair of events in terms of taxonomic categories. And then subtract the value returned by $SimEvents(e_t, e'_{t'})$ from one to obtain the diversity measure for a pair of events. For illustration, consider a pair of sets of taxonomic categories $e_t = t_0, t_1, t_2$ and $e'_{t'} = t'_0, t'_1, t'_2$ as shown in Table 6.2

We start with the procedure $SimEvents$ to compute the similarity between a pair of

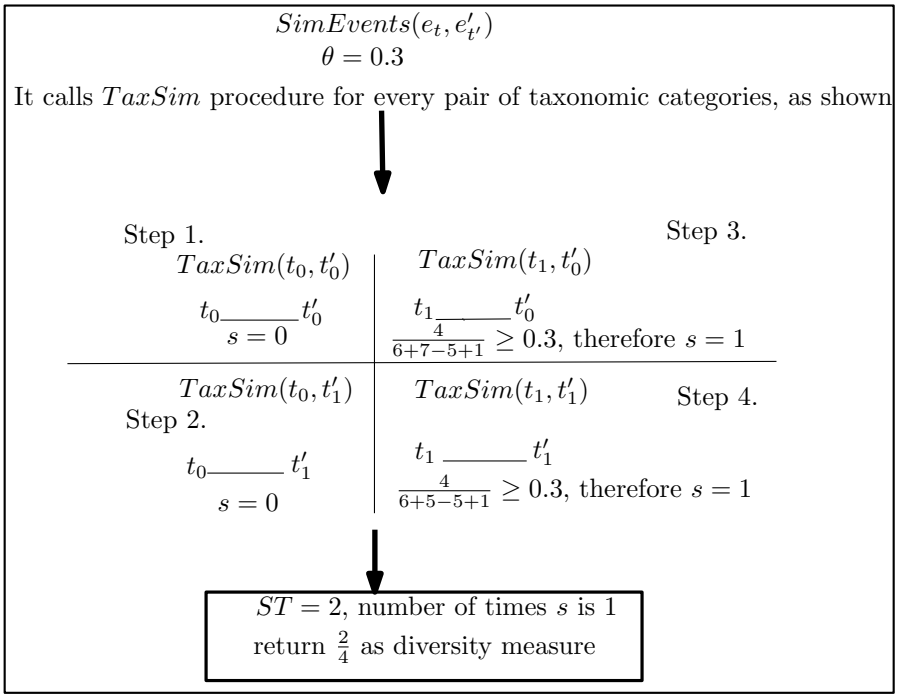


FIGURE 6.1: Illustration: To compute the diversity measure for a pair of events on basis of taxonomic categories

events e and e', and then it calls the procedure $TaxSim$ for every pair of taxonomic categories as shown in Figure 6.1. Then the value of ST is computed, which is the count of number of times s equals to 1 divided by the sum of the set sizes of taxonomic categories of events e and e'.

6.4 Conclusion

The diversity measures can be modeled for taxonomic categories, terms or tags; depending on the topic or event model. For example, if our topics are defined as set of taxonomic categories then we can find the diversity measures using algorithms 1 and 2, and if documents inside an event are identified as unique ids, terms, or tags then we can use the Jaccard Index [40] for finding diversity for a pair of events.

Chapter 7

Experiments and Evaluation

7.1 Experimental Setup

The New York Times archive dataset is used for the experiments. This dataset consists of news articles published between 1987 and 2007. It comprises of 1.8 million documents and 84.84% documents are annotated with keyword descriptors. We use these descriptors as tags. For experiment purposes, we detect events at the timelines specified in Table 7.1 and then apply personalization and diversification model for a user specified interest to the events detected at the respective timeline. The implementation of the retrieval models for personalizing and the greedy solution of diversity methods is done in Java 1.6 on Eclipse IDE. Further, to support the linear programming implementation of diversity problems, java libraries provided by IBM ILOG CPLEX 12.2 are used [1].

User Interest	**Timeline**
Catastrophe	Year 2004, Month 12, Week 4.
Stock Market	Year 1994, Month 7, Week 3.
Computer Science	Year 2001, Month 6, Week 3.
Health and Medicine	Year 2003, Month 10, Week 3.
Military Affairs	Year 2001, Month 12, Week 2.
English Books	Year 2000, Month 6, Week 3.
Space Science	Year 2005, Month 7, Week 4.
Europe Politics	Year 1992, Month 9, Week 3.
Environment Hazards	Year 2003, Month 6, Week 3.

TABLE 7.1: List of user interests and the respective timelines

The evaluation is conducted by 5 computer scientists. In the current evaluation, we refer them as User-A, User-B, User-C, User-D, and User-E.

[1] Thanks to the IBM Academic Initiative Program for providing the free academic license of IBM ILOG CPLEX Optimizer

7.2 Evaluation

Our experimental study of the personalization and diversification model is based on the following evaluations.

1. Quality Evaluation : We compare the quality of results produced by the retrieval models, presented as ESD-LM, ESD-TFIDF, EHD-LM, and EHD-TFIDF, based on evaluations by users.
2. Performance Evaluation : Firstly, we compare the solution obtained by greedy for Maximum Diversity Quality Problem with the solution obtained through CPLEX 12.2 linear programming solver. Secondly, we compared the running time of CPLEX 12.2 with the greedy solution.

7.2.1 Quality Evaluation

Our evaluation measures are based on the following assumptions.

1. Given a user's information need, represented by keywords, each event in a given collection of events is either relevant or nonrelevant with respect to the specified user interest as keywords.
2. The relevance of an event e depends only on the user information need and the e itself. It is independent of the ranking of the events according to a user interest. It means that a user may have found an event ranked at position, say 3^{rd}, as relevant and an event ranked at position, say 2^{nd}, as irrelevant according to his interest.

On the basis of above assumptions, we proposed the following evaluation measures.

7.2.1.1 Precision@k

This measure models the satisfaction of a user for each interest specified as keywords by presenting with a list of up to k top ranked personalized detected events. It is defined as

$$Precision@k = \frac{|Results[1...k] \cap Relevant|}{k},$$

where $Results[1...k]$ consists of the Top-k events returned by the retrieval model and $Relevant$ is the set of relevant events. In our problem setting we have set k to 5 and 10. In order to measure $Precision@k$, we asked the users to inspect the k events which were retrieved by the models for 8 different pre-defined user interests. They could mark

them as relevant or irrelevant by analyzing the information (as documents) related with each event. Thereafter, we divide the count of the number of relevant events according to a user by k to obtain $Precision@k$. The precision values obtained for 8 different user interests and for all retrieval approaches from the analyzes of results by 5 users is given in tables 7.2 ($Precision@5$) and 7.3 ($Precision@10$). The higher the precision, higher the satisfiability of a user with respect to his interest.

Considering $Precision@5$ shown in Table 7.2, first, it is observed for all user interests the results obtained by ESD-LM and ESD-TFIDF are equivalent and also the results obtained by EHD-LM and EHD-TFIDF are equivalent. Second, for the user interests *Stock Market*, *Catastrophe*, *Computer Science, and English Books* (marked in blue color), users are equally satisfied with the results obtained by all retrieval models. Hence, it is not possible to say which model is best for these user interests. Whereas for the user interests *Health and Medicine, Europe Politics, and Military Affairs*, among the ESD and ESD models, there is no clear winner. It can be noticed that User-C equally prefers the results obtained from ESD and EHD models for a user interest *Military Affairs*(marked in red color). Also, for the user interest *Health and Medicine*, User-D equally prefers (with only 0.2 precision) ESD and EHD models (marked in green color). Hence from this table it is not very clear that the results of which model satisfies the users. In order to obtain better estimates of precision, we evaluate the same user interests for precision at 10.

Considering $Precision@10$ shown in Table 7.3, first, it is observed the top-10 results do not completely overlap for all retrieval models. Therefore these results can be indirectly used for inferring which models best satisfies the users for k 10. Second, for a user interest *Catastrophe*, the precision values are maximum for ESD-LM and precision is 0 (worst) for EHD-LM and EHD-TFIDF (marked in blue color). Third, for a user interest *Stock Market*, the precision values are the same as the results obtained by all models are equivalent (marked in red color, see AppendixA).

7.2.1.2 MacroPrecision@k

To explore which retrieval model dealt best with every user interest, the macro precision is observed. This measure models the satisfaction of a user for a set of interests specified as keywords. For a set of user interests U, it is defined as

$$MacroPrecision@k = \frac{1}{|U|} \sum_{u \in U} Precision@k(u)$$

The macro precision values obtained for all retrieval approaches from the evaluation of results by 5 users is given in the figures 7.1 ($MacroPrecision@5$) and 7.2 ($MacroPrecision@10$).

Event as Set of Documents - Language Model (ESD-LM) and, Event as a Set of Documents - TFIDF (ESD-TFIDF)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	1.0	1.0	1.0	1.0	1.0
Military Affairs	0.8	0.8	0.4	0.6	1.0
Health and Medicine	0.4	0.6	0.8	0.2	0.6
Catastrophe	1.0	0.8	0.4	1.0	1.0
Computer Science	0.4	0.8	0.4	0.8	0.8
English Books	0.8	0.4	0.4	0.8	0.6
Space Science	0.2	1.0	1.0	0.8	0.8
Europe Politics	1.0	1.0	1.0	1.0	1.0

Event as a Huge Document - Language Model (EHD-LM) and, Event as a Huge Document - TFIDF (EHD-TFIDF)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	1.0	1.0	1.0	1.0	1.0
Military Affairs	0.6	0.6	0.4	0.4	0.8
Health and Medicine	0.2	0.4	0.6	0.2	0.4
Catastrophe	1.0	0.8	0.4	1.0	1.0
Computer Science	0.4	0.8	0.4	0.8	0.8
English Books	0.8	0.4	0.4	0.8	0.6
Space Science	0	1.0	1.0	0.6	0.2
Europe Politics	0.8	0.8	0.8	0.8	0.8

TABLE 7.2: Precision@5

For *MacroPrecision*@5 (Refer Figure 7.1), *User A* is 62% satisfied with the results retrieved by ESD-LM and ESD-TFIDF. Approximately 53% satisfied with the results retrieved by EHD-LM and EHD-TFIDF. And *User B* is 73% satisfied with the results obtained from ESD-LM and ESD-TFIDF and is least satisfied with the results obtained from the EHD-LM and EHD-TFIDF model. The similar scenario is observed from the macro precision plots of other users, too. The common observation is that all users prefer the results obtained by ESD-LM and ESD-TFIDF. *Hence, it is concluded that user is overall much more satisfied with retrieval results obtained by the ESD-LM and ESD-TFIDF.*

For*MacroPrecision*@10 (Refer Figure 7.2), *User A, User B, User C, User D and User E* are maximum satisfied with the results obtained by ESD-LM. Their satisfaction percentages are 47%, 56%, 48%, 53% and 58%, respectively. This result give an impression that our ESD-LM is able to satisfy on average 52.4% to users. Moreover, the retrieved results are dependent on the test data set, so the richness of the data set will also effect the results.

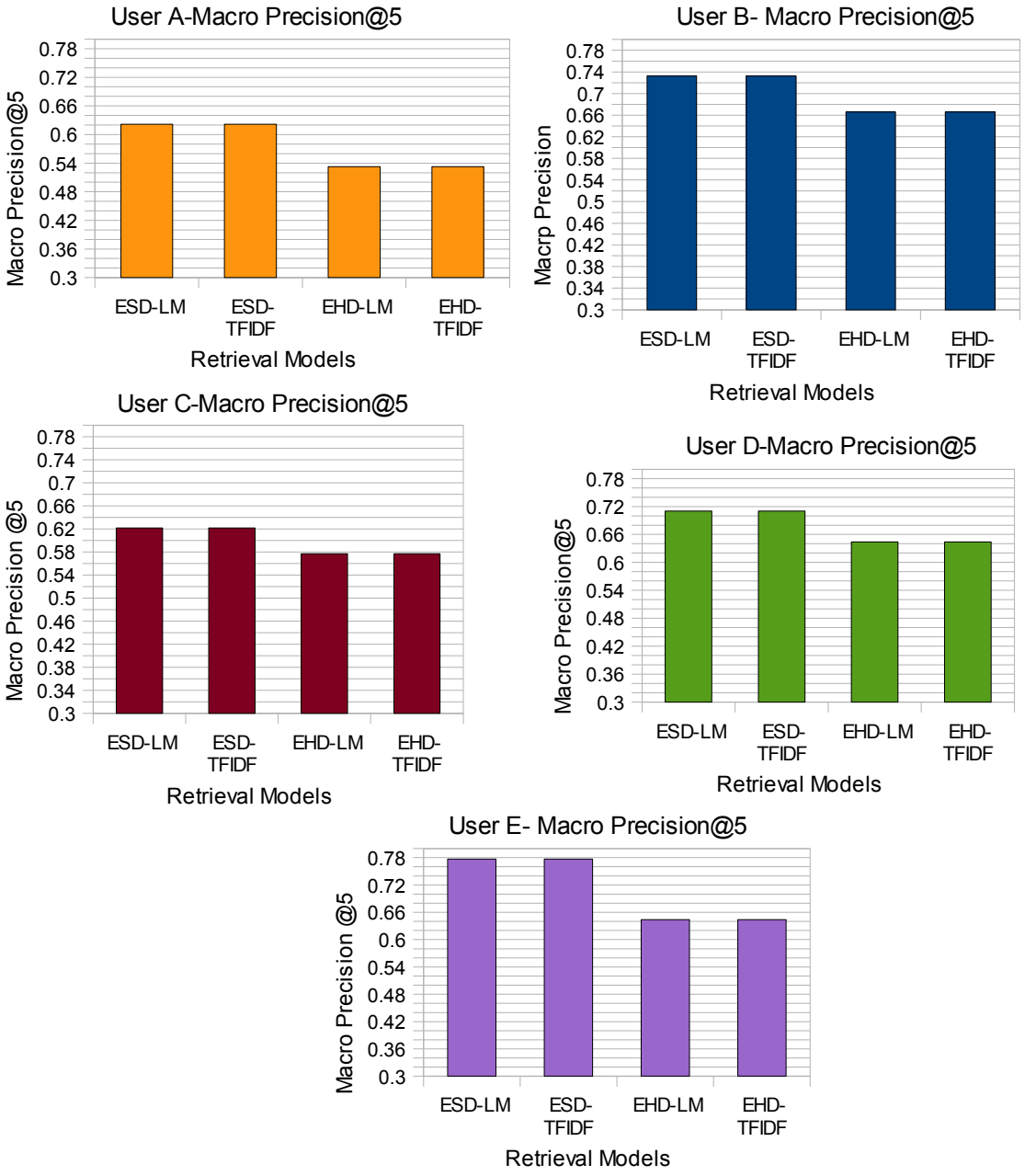


FIGURE 7.1: MacroPrecision@5 Vs Retrieval Models

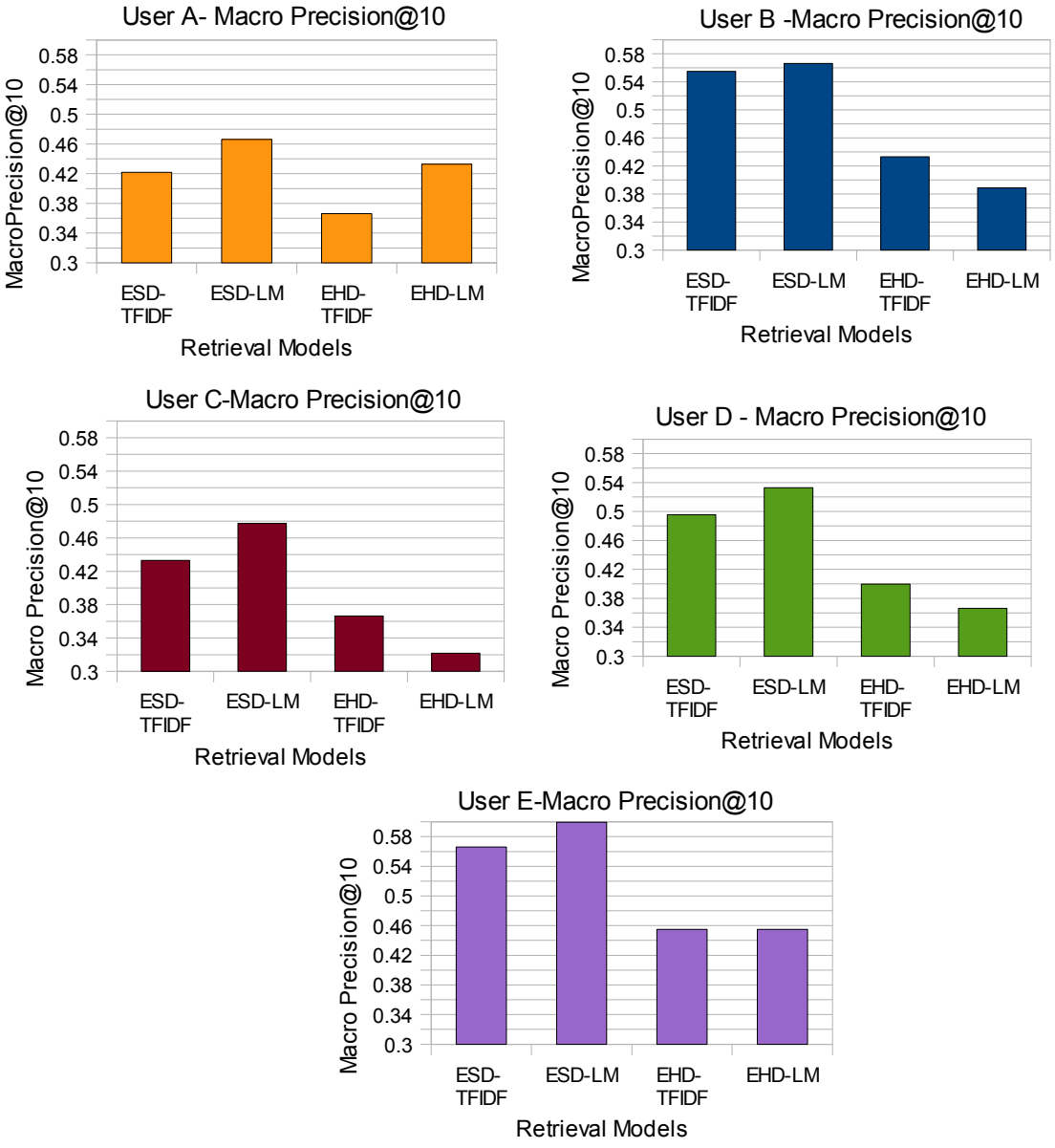


FIGURE 7.2: MacroPrecision@10 Vs Retrieval Models

Event as Set of Documents- Language Model (ESD-LM)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	0.8	0.7	0.7	0.7	0.8
Military Affairs	0.5	0.4	0.2	0.4	0.6
Health and Medicine	0.5	0.7	0.7	0.6	0.5
Catastrophe	0.8	0.8	0.5	1.0	0.8
Computer Science	0.3	0.5	0.3	0.4	0.5
English Books	0.4	0.2	0.2	0.5	0.6
Space Science	0.2	0.9	1.0	0.5	0.7
Europe Politics	0.7	0.8	0.6	0.6	0.8

Event as Set of Documents -TFIDF (ESD-TFIDF)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	0.8	0.7	0.7	0.7	0.8
Military Affairs	0.5	0.4	0.2	0.4	0.6
Health and Medicine	0.3	0.6	0.4	0.4	0.4
Catastrophe	0.6	0.8	0.4	0.9	0.6
Computer Science	0.3	0.5	0.3	0.4	0.5
English Books	0.4	0.2	0.2	0.5	0.6
Space Science	0.2	0.9	1.0	0.5	0.7
Europe Politics	0.7	0.8	0.6	0.6	0.8

Event as a Huge Document -Language Model (EHD-LM)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	0.8	0.7	0.7	0.7	0.8
Military Affairs	0.5	0.4	0.2	0.4	0.6
Health and Medicine	0.3	0.3	0.3	0.3	0.3
Catastrophe	0	0	0	0	0
Computer Science	0.2	0.5	0.3	0.4	0.5
English Books	0.4	0.2	0.2	0.5	0.6
Space Science	1.0	0.5	0.5	0.3	0.4
Europe Politics	0.7	0.8	0.6	0.6	0.8

Event as a Huge Document - TFIDF (EHD-TFIDF)

User Interest	User-A	User-B	User-C	User-D	User-E
Stock Market	0.8	0.7	0.7	0.7	0.8
Military Affairs	0.5	0.5	0.2	0.4	0.5
Health and Medicine	0.6	0.6	0.7	0.6	0.4
Catastrophe	0	0	0	0	0
Computer Science	0.2	0.5	0.3	0.4	0.5
English Books	0.4	0.2	0.2	0.5	0.6
Space Science	0.1	0.5	0.5	0.3	0.4
Europe Politics	0.7	0.8	0.6	0.6	0.8

TABLE 7.3: Precision@10

7.2.2 Performance Evaluation

Our performance based evaluation for the Maximum Diversity Quality Problem (MDQP) is based on following measures:

Parameters : n=10,m=5

User Interest	$Solution_{CPLEX} \cap Solution_{greedy}$	%Accuracy
Stock Market	4	80
Military Affairs	5	100
Health and Medicine	3	60
Catastrophe	4	80
Computer Science	4	80
English Books	5	100
Space Science	4	80
Europe Politics	3	60

TABLE 7.4: % Accuracy of the Greedy Solution for MDQP

7.2.2.1 %Accuracy of the Greedy solution

In this evaluation measure, we aim to compare the solution obtained by the greedy method to the solution obtained by CPLEX12.2 linear programming solver. This is done for the performance evaluation of the solution obtained by the greedy. For a given set of n events, the MDQP consists of determining a subset, say m events, such that the sum of the product of diversity measure and quality measure of events is maximum. The % Accuracy is measured as,

$$\%Accuracy = \frac{Solution_{CPLEX} \cap Solution_{greedy}}{m} * 100 \tag{7.1}$$

where $Solution_{CPLEX}$ and $Solution_{greedy}$ are the solution sets obtained through CPLEX and greedy method, respectively. In our experimental setting n is 10 and m is 5 and results are given in the Table 7.4.

It is observed that the solution obtained by greedy for Maximum Diversity Quality Problem is not so far from the optimal solution obtained through CPLEX12.2 linear programming solver. Moreover for the user interests, *Military Affairs* and *English Books*, the greedy method returned the optimal solution.

7.2.2.2 Running time of CPLEX12.2 VS Greedy solution

Though the attainment of exact solutions for the diversity problem rely on the LP solvers, solving a large sized problem with the exact method takes an excessively long time. For small instances, the solution for the diversity problem can be attained in negligible time through greedy solution as compared to the one obtained by CPLEX12.2. In order to show the differences in the running time, we explored it for small instance (n=10, m=5) and found that the greedy returns the solution in less than .01 milliseconds for all user interests. Whereas, the LP solver takes a considerably long time in comparison to

User Interest	**CPLEX12.2 (msecs)**	**Greedy (msecs)**
Stock Market	276	
Military Affairs	199	
Health and Medicine	413	greedy returns
Catastrophe	343	in less than
Computer Science	461	.01 for all
English Books	187	user interests
Space Science	493	
Europe Politics	532	

TABLE 7.5: Running time of CPLEX12.2

greedy. It is clear that greedy solution vastly outperforms CPLEX12.2 in terms of the running time for the Maximum Diversity Quality Problem. As a part of future work, we propose the following extensions of the experimental study

1. Incorporating the feedback of more users to evaluate the results retrieved by the methods (ESD-LM, ESD-TFIDF, EHD-LM, and EHD-TFIDF).

2. Improving the personalized results by user expansion techniques and diversity methods and then get the results evaluated by users. This study can help us in knowing that how much expansion and diversification helps in improving the personalized results.

3. Comparing the accuracy of the greedy methods with linear programming solution obtained through CPLEX for both MDQP and MMDQP and for various instances.

4. Running experiments for different data sets like tweets and blogs.

Chapter 8

Conclusion

We developed a *Personalization and Diversification Framework* for the events detected by online social media to satisfy the information needs of users. Our proposed framework is composed is of three different modules, i.e, Retrieval, Expansion, and Diversification. The Retrieval module is responsible for retrieving events according to user's interest. The Expansion module refines the ranking of the retrieved results. And, the Diversification module aims to eliminate redundant events from the result set such that there is a diversity in the Top-k events presented to a user. We briefly tried to detail an experimental study and we obtain the following conclusions : First, the retrieval results produced by ESD-LM better satisfy the users in comparison to other retrieval methods. Second, the solution obtained by greedy for Maximum Diversity Quality Problem is not so far from the optimal solution obtained through linear programming solver. Third, the greedy method vastly outperforms linear programming solver in terms of the running time for the Maximum Diversity Quality Problem.

In our present work, we are personalizing the events according to the user specified interest as keywords, but, in the future, we can extend or work to detect user interests by tracking their spatial information. The one possible way of obtaining spatial information is by tracking the user's profile. Moreover, we can extend our work for personalizing streams of emerging topics, or personalizing news stories.

Appendix A

Experiment Results - Retrieval Models

User Interest: Space Science

Year:2005, Month:07, Week:4

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
Terrorism**Bombs and Explosives	Space Shuttle**Discovery (Space Shuttle)	International Space Station**Space Shuttle	International Space Station**Space Shuttle	International Space Station**Space Shuttle
United States International Relations**United States Armament and Defense	Space**Discovery (Space Shuttle)	International Space Station**Space	International Space Station**Space	International Space Station**Space
Terrorism**Transit Systems	Space**Space Shuttle	International Space Station**Discovery (Space Shuttle)	International Space Station**Discovery (Space Shuttle)	International Space Station**Discovery (Space Shuttle)
Transit Systems**Bombs and Explosives	Space**Accidents and Safety	International Space Station**Accidents and Safety	International Space Station**Accidents and Safety	International Space Station**Accidents and Safety
Subways**Terrorism	Accidents and Safety**Discovery (Space Shuttle)	Columbia (Space Shuttle)**International Space Station	Columbia (Space Shuttle)**International Space Station	Columbia (Space Shuttle)**International Space Station
Subways**Transit Systems	Accidents and Safety**Space Shuttle	Foam**Discovery (Space Shuttle)	Space**Accidents and Safety	Space**Accidents and Safety
Subways**Buses	Columbia (Space Shuttle)**Discovery (Space Shuttle)	International Space Station**Foam	Accidents and Safety**Discovery (Space Shuttle)	Accidents and Safety**Discovery (Space Shuttle)
Terrorism**Buses	Columbia (Space Shuttle)**Space Shuttle	Columbia (Space Shuttle)**Foam	Accidents and Safety**Space Shuttle	Accidents and Safety**Space Shuttle
Terrorism**United States International Relations	Columbia (Space Shuttle)**Space	Columbia (Space Shuttle)**Discovery (Space Shuttle)	Columbia (Space Shuttle)**Discovery (Space Shuttle)	Columbia (Space Shuttle)**Discovery (Space Shuttle)
Space Shuttle**Discovery (Space Shuttle)	Columbia (Space Shuttle)**Accidents and Safety	Columbia (Space Shuttle)**Space Shuttle	Columbia (Space Shuttle)**Space Shuttle	Columbia (Space Shuttle)**Space Shuttle

User Interest: Military Affairs

Year:2001, Month:12, Week:2

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
Hijacking**Airlines and Airplanes	Airlines and Airplanes**Military Aircraft	Airlines and Airplanes**Military Aircraft	Airlines and Airplanes**Military Aircraft	Airlines and Airplanes**Military Aircraft
World Trade Center (NYC)**Airlines and Airplanes	Two Thousand and One (Year)**United States Armament and Defense	Two Thousand and One (Year)**United States Armament and Defense	Politics and Government**Legislatures and Parliaments	Politics and Government**Legislatures and Parliaments
World Trade Center (NYC)**Hijacking	Politics and Government**Legislatures and Parliaments	Politics and Government**Legislatures and Parliaments	Two Thousand and One (Year)**United States Armament and Defense	Two Thousand and One (Year)**United States Armament and Defense
Terrorism**Airlines and Airplanes	Terrorism**Two Thousand and One (Year)	Terrorism**Two Thousand and One (Year)	Accidents and Safety**Airlines and Airplanes	Accidents and Safety**Airlines and Airplanes
World Trade Center (NYC)**Terrorism	Terrorism**Accidents and Safety	Terrorism**Accidents and Safety	Terrorism**Two Thousand and One (Year)	Terrorism**Two Thousand and One (Year)
Terrorism**Hijacking	Accidents and Safety**Airlines and Airplanes	Accidents and Safety**Airlines and Airplanes	Terrorism**Accidents and Safety	Terrorism**Accidents and Safety
World Trade Center (NYC)**United States Armament and Defense	Two Thousand and One (Year)**Pentagon Building	Two Thousand and One (Year)**Pentagon Building	Budgets and Budgeting**Taxation	Budgets and Budgeting**Taxation
Airlines and Airplanes**United States Armament and Defense	World Trade Center (NYC)**Two Thousand and One (Year)	World Trade Center (NYC)**Two Thousand and One (Year)	Budgets and Budgeting**Airlines and Airplanes	Budgets and Budgeting**Airlines and Airplanes
Hijacking**United States Armament and Defense	Hijacking**Two Thousand and One (Year)	Hijacking**Two Thousand and One (Year)	Budgets and Budgeting**Terrorism	Budgets and Budgeting**Terrorism
Terrorism**United States Armament and Defense	Budgets and Budgeting**Economic Conditions and Trends	Budgets and Budgeting**Economic Conditions and Trends	Immigration and Refugees**United States International Relations	Immigration and Refugees**United States International Relations

User Interest: Health and Medicine

Year:2003, Month:10, Week:3

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
United States International Relations**United States Armament and Defense	Medicine and Health**Doctors	Medicine and Health**Health Insurance and Managed Care	Medicine and Health**Health Insurance and Managed Care	Medicine and Health**Health Insurance and Managed Care
Terrorism**United States International Relations	Medicine and Health**Health Insurance and Managed Care	Murders and Attempted Murders**Children and Youth	Murders and Attempted Murders**Children and Youth	Aged**Health Insurance and Managed Care
Terrorism**United States Armament and Defense	Medicine and Health**Drugs (Pharmaceuticals)	Medicine and Health**Computers and the Internet	Computers and the Internet**International Trade and World Market	Medicine and Health**Drugs (Pharmaceuticals)
Baseball**Playoff Games	Medicine and Health**Computers and the Internet	Computers and the Internet**International Trade and World Market	Medicine and Health**Computers and the Internet	Medicine and Health**Doctors
Baseball**World Series	Murders and Attempted Murders**Children and Youth	Medicine and Health**Drugs (Pharmaceuticals)	Medicine and Health**Drugs (Pharmaceuticals)	Medicine and Health**Aged
Budgets and Budgeting**United States International Relations	Aged**Health Insurance and Managed Care	Medicine and Health**Doctors	Computers and the Internet**Advertising and Marketing	Medicare**Aged
Election Issues**Presidential Election of 2004	Computers and the Internet**International Trade and World Market	Aged**Health Insurance and Managed Care	Aged**Health Insurance and Managed Care	Medicine and Health**Medicare
Accidents and Safety**Ferries	Architecture**Area Planning and Renewal	Hotels and Motels**Travel and Vacations	Medicine and Health**Doctors	Medicine and Health**Computers and the Internet
Politics and Government**United States International Relations	Medicine and Health**Aged	Computers and the Internet**Children and Youth	Tests and Testing**Education and Schools	Computers and the Internet**International Trade and World Market
Budgets and Budgeting**Terrorism	Medicare**Aged	Computers and the Internet**Advertising and Marketing	Demonstrations and Riots**Politics and Government	Housing**Interior Design

User Interest: English Books

Year:2000, Month:07, Week:2

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
Election Issues**Presidential Election of 2000	English Language**Books and Literature	English Language**Books and Literature	English Language**Books and Literature	English Language**Books and Literature
Telephones and Telecommunications**Cellular Telephones	Computers and the Internet**Books and Literature	Children and Youth**Books and Literature	Children and Youth**Books and Literature	Children and Youth**Books and Literature
Police Brutality and Misconduct**Police	Children and Youth**Books and Literature	Computers and the Internet**Books and Literature	Computers and the Internet**Books and Literature	Computers and the Internet**Books and Literature
Palestinians**United States International Relations	Cooking and Cookbooks**Recipes	Cooking and Cookbooks**Recipes	Cooking and Cookbooks**Recipes	Cooking and Cookbooks**Recipes
Wimbledon Tennis Tournament**Tennis	Books and Literature**Book Trade	Books and Literature**Book Trade	Books and Literature**Book Trade	Books and Literature**Book Trade
Baseball**All-Star Games	Barbecue**Recipes	Barbecue**Recipes	Barbecue**Recipes	Barbecue**Recipes
Stocks and Bonds**Mutual Funds	Computers and the Internet**Travel and Vacations	Computers and the Internet**Travel and Vacations	Computers and the Internet**Travel and Vacations	Computers and the Internet**Travel and Vacations
Medicine and Health**Drugs (Pharmaceuticals)	Advertising**Computers and the Internet	Advertising**Computers and the Internet	Advertising**Computers and the Internet	Colleges and Universities**Education and Schools
Recordings (Audio)**Music	Travel and Vacations**Restaurants	Travel and Vacations**Restaurants	Colleges and Universities**Education and Schools	Advertising**Computers and the Internet
Telephones and Telecommunications**Mergers, Acquisitions and Divestitures	Tests and Testing**Colleges and Universities	Tests and Testing**Colleges and Universities	Travel and Vacations**Restaurants	Tests and Testing**Colleges and Universities

User Interest: Europe Politics

Year:1992, Month:09, Week:3

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
ELECTION ISSUES**PRESIDENTIAL ELECTION OF 1992	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT
PRIMARIES**ELECTIONS	EUROPEAN MONETARY SYSTEM**POLITICS AND GOVERNMENT	EUROPEAN MONETARY SYSTEM**POLITICS AND GOVERNMENT	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**EUROPEAN MONETARY SYSTEM	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**EUROPEAN MONETARY SYSTEM
HURRICANE ANDREW**HURRICANES AND TROPICAL STORMS	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**EUROPEAN MONETARY SYSTEM	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**EUROPEAN MONETARY SYSTEM	EUROPEAN MONETARY SYSTEM**ELECTIONS	EUROPEAN MONETARY SYSTEM**ELECTIONS
CONCERTS AND RECITALS**MUSIC	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**ELECTIONS	EUROPEAN MONETARY SYSTEM**ELECTIONS	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**ELECTIONS	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**ELECTIONS
ART**ART SHOWS	POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**ELECTIONS	EUROPEAN MONETARY SYSTEM**POLITICS AND GOVERNMENT	EUROPEAN MONETARY SYSTEM**POLITICS AND GOVERNMENT
COLLEGE ATHLETICS**FOOTBALL	EUROPEAN MONETARY SYSTEM**ELECTIONS	POLITICS AND GOVERNMENT**TREATIES	POLITICS AND GOVERNMENT**TREATIES	POLITICS AND GOVERNMENT**TREATIES
ELECTION RESULTS**ELECTIONS	ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT	POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS	POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS	EUROPEAN MONETARY SYSTEM**ECONOMIC CONDITIONS AND TRENDS
POLICE**DEMONSTRATIONS AND RIOTS	POLITICS AND GOVERNMENT**CIVIL WAR AND GUERRILLA WARFARE	ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT	ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT	EUROPEAN MONETARY SYSTEM**CREDIT
DRAFT AND RECRUITMENT (MILITARY)**PRESIDENTIAL ELECTION OF 1992	POLITICS AND GOVERNMENT**TREATIES	EUROPEAN MONETARY SYSTEM**CREDIT	EUROPEAN MONETARY SYSTEM**CREDIT	POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS
MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**EUROPEAN MONETARY SYSTEM	EUROPEAN MONETARY SYSTEM**CREDIT	POLITICS AND GOVERNMENT**CIVIL WAR AND GUERRILLA WARFARE	EUROPEAN MONETARY SYSTEM**ECONOMIC CONDITIONS AND TRENDS	ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT

User Interest: Catastrophe

Year:2004, Month:12, Week:4

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
United States International Relations**United States Armament and Defense	Medicine and Health**Earthquakes	Philanthropy**Tidal Waves	Governors (US)**Election Results	Governors (US)**Election Results
Tidal Waves**Earthquakes	Medicine and Health**Water	Philanthropy**Earthquakes	Vaccination and Immunization**Influenza	Vaccination and Immunization**Influenza
Politics and Government**Elections	Medicine and Health**Tidal Waves	Medicine and Health**Earthquakes	Travel and Vacations**Skiing	Travel and Vacations**Skiing
Terrorism**United States Armament and Defense	Philanthropy**Tidal Waves	Medicine and Health**Water	Ratings and Rating Systems**Recordings (Audio)	Ratings and Rating Systems**Recordings (Audio)
Terrorism**United States International Relations	Philanthropy**Earthquakes	Medicine and Health**Tidal Waves	Frauds and Swindling**Politics and Government	Frauds and Swindling**Politics and Government
Foreign Aid**Earthquakes	Water**Tidal Waves	Foreign Aid**Philanthropy	Condominiums**Housing	Condominiums**Housing
Foreign Aid**Tidal Waves	Water**Earthquakes	Water**Tidal Waves	Politics and Government**United States International Relations	Politics and Government**United States International Relations
United States International Relations**Bombs and Explosives	Foreign Aid**Earthquakes	Water**Earthquakes	Christmas**United States Armament and Defense	Christmas**United States Armament and Defense
United States Armament and Defense**Bombs and Explosives	Foreign Aid**Tidal Waves	Foreign Aid**Earthquakes	Politics and Government**Legislatures and Parliaments	Politics and Government**Legislatures and Parliaments
Music**Recordings (Audio)	Earthquakes**Tsunamis	Foreign Aid**Tidal Waves	Music**Computers and the Internet	Music**Computers and the

User Interest: Computer Science

Year:2001, Month:06, Week:3

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
Election Issues**Elections	Computers and the Internet**Mergers, Acquisitions and Divestitures	Computers and the Internet**Computer Chips	Computer Security**Computers and the Internet	Computer Security**Computers and the Internet
United States Open (Golf)**Golf	Computers and the Internet**Books and Literature	Computer Security**Computers and the Internet	Computers and the Internet**Computer Chips	Computers and the Internet**Computer Chips
Primaries**Elections	Computer Security**Computers and the Internet	Computers and the Internet**Computer Software	Computers and the Internet**Computer Software	Computers and the Internet**Computer Software
Medicine and Health**Health Insurance	Computers and the Internet**Stocks and Bonds	Computers and the Internet**Mergers, Acquisitions and Divestitures	Computers and the Internet**Mergers, Acquisitions and Divestitures	Computers and the Internet**Mergers, Acquisitions and Divestitures
Tests and Testing**Education and Schools	Computers and the Internet**Computer Chips	Computers and the Internet**Books and Literature	Computers and the Internet**Books and Literature	Computers and the Internet**Books and Literature
Computers and the Internet**Computer Software	Computers and the Internet**Computer Software	Bankruptcies**Computers and the Internet	Bankruptcies**Computers and the Internet	Bankruptcies**Computers and the Internet
United States International Relations**United States Armament and Defense	Computers and the Internet**Children and Youth	Computers and the Internet**Stocks and Bonds	Computers and the Internet**Stocks and Bonds	Computers and the Internet**Stocks and Bonds
Governors (US)**Elections	Telephones and Telecommunications**Computers and the Internet	Computers and the Internet**Music	Computers and the Internet**Music	Telephones and Telecommunications**Computers and the Internet
Missiles and Missile Defense Systems**United States Armament and Defense	Fiber Optics**Computers and the Internet	Computers and the Internet**Children and Youth	Computers and the Internet**Layoffs and Job Reductions	Computers and the Internet**Music
Explosions**Fires and Firemen	Bankruptcies**Computers and the Internet	Fiber Optics**Computers and the Internet	Fiber Optics**Computers and the Internet	Computers and the Internet**Layoffs and Job Reductions

User Interest: Stock Market

Year:1994, Month:07, Week:4

No Personalization	ESD-LM	ESD-TFIDF	EHD-LM	EHD-TFIDF
MEDICINE AND HEALTH**HEALTH INSURANCE	STOCKS AND BONDS**DOW JONES STOCK AVERAGE	STOCKS AND BONDS**DOW JONES STOCK AVERAGE	STOCKS AND BONDS**DOW JONES STOCK AVERAGE	STOCKS AND BONDS**DOW JONES STOCK AVERAGE
REFORM AND REORGANIZATION**MEDICINE AND HEALTH	MUTUAL FUNDS**STOCKS AND BONDS	BROKERS AND BROKERAGE FIRMS**STOCKS AND BONDS	BROKERS AND BROKERAGE FIRMS**STOCKS AND BONDS	BROKERS AND BROKERAGE FIRMS**STOCKS AND BONDS
REFORM AND REORGANIZATION**HEALTH INSURANCE	BROKERS AND BROKERAGE FIRMS**STOCKS AND BONDS	MUTUAL FUNDS**STOCKS AND BONDS	MUTUAL FUNDS**STOCKS AND BONDS	MUTUAL FUNDS**STOCKS AND BONDS
NYTRAVEL**TRAVEL AND VACATIONS	DERIVATIVES (FINANCIAL TRANSACTIONS)**STOCKS AND BONDS	DERIVATIVES (FINANCIAL TRANSACTIONS)**STOCKS AND BONDS	DERIVATIVES (FINANCIAL TRANSACTIONS)**STOCKS AND BONDS	LAW AND LEGISLATION**INTERNATIONAL TRADE AND WORLD MARKET
LAW AND LEGISLATION**REFORM AND REORGANIZATION	INVESTMENT STRATEGIES**STOCKS AND BONDS	INVESTMENT STRATEGIES**MUTUAL FUNDS	INVESTMENT STRATEGIES**STOCKS AND BONDS	DERIVATIVES (FINANCIAL TRANSACTIONS)**STOCKS AND BONDS
LAW AND LEGISLATION**HEALTH INSURANCE	INVESTMENT STRATEGIES**MUTUAL FUNDS	INVESTMENT STRATEGIES**STOCKS AND BONDS	INVESTMENT STRATEGIES**MUTUAL FUNDS	INVESTMENT STRATEGIES**STOCKS AND BONDS
LAW AND LEGISLATION**MEDICINE AND HEALTH	STOCKS AND BONDS**VIOLATIONS OF SECURITIES AND COMMODITIES REGULATIONS	STOCKS AND BONDS**VIOLATIONS OF SECURITIES AND COMMODITIES REGULATIONS	STOCKS AND BONDS**VIOLATIONS OF SECURITIES AND COMMODITIES REGULATIONS	INVESTMENT STRATEGIES**MUTUAL FUNDS
UNITED STATES INTERNATIONAL RELATIONS**UNITED STATES ARMAMENT AND DEFENSE	STOCKS AND BONDS**CREDIT	STOCKS AND BONDS**CREDIT	STOCKS AND BONDS**CREDIT	STOCKS AND BONDS**CREDIT
CIVIL WAR AND GUERRILLA WARFARE**IMMIGRATION AND REFUGEES	LAW AND LEGISLATION**INTERNATIONAL TRADE AND WORLD MARKET	LAW AND LEGISLATION**INTERNATIONAL TRADE AND WORLD MARKET	LAW AND LEGISLATION**INTERNATIONAL TRADE AND WORLD MARKET	STOCKS AND BONDS**VIOLATIONS OF SECURITIES AND COMMODITIES REGULATIONS
CHOLERA**IMMIGRATION AND REFUGEES	UNITED STATES ECONOMY**CREDIT	UNITED STATES ECONOMY**CREDIT	UNITED STATES ECONOMY**CREDIT	UNITED STATES ECONOMY**CREDIT

Appendix B

Diversity results

User Interest – Catastrophe Parameters n=10, m=5	
CPLEX 12.2	**GREEDY SOLUTION**
Medicine and Health**Earthquakes	Medicine and Health**Earthquakes
Medicine and Health**Water	Medicine and Health**Water
Philanthropy**Earthquakes	Medicine and Health**Tidal Waves
Philanthropy**Tidal Waves	Philanthropy**Earthquakes
Earthquakes**Tsunamis	Philanthropy**Tidal Waves

User Interest – Health and Medicine Parameters n=10, m=5	
CPLEX 12.2	**GREEDY SOLUTION**
Medicine and Health**Doctors	Medicine and Health**Doctors
Medicine and Health**Health Insurance and Managed Care	Medicine and Health**Health Insurance and Managed Care
Medicine and Health**Computers And the Internet	Medicine and Health**Drugs (Pharmaceuticals)
Murders and Attempted Murders**Children And Youth	Medicine and Health**Computers And the Internet
Architecture**Area Planning And Renewal	Computers and the Internet**International Trade and World Market

User Interest -Europe Politics Parameters n=10, m=5	
CPLEX12.2	**GREEDY SOLUTION**
MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT	MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**POLITICS AND GOVERNMENT
MAASTRICHT TREATY ON EUROPEAN POLITICAL AND MONETARY UNION**ELECTIONS	EUROPEAN MONETARY SYSTEM**POLITICS AND GOVERNMENT
POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS	POLITICS AND GOVERNMENT**ECONOMIC CONDITIONS AND TRENDS
ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT	ARMAMENT, DEFENSE AND MILITARY FORCES**POLITICS AND GOVERNMENT
EUROPEAN MONETARY SYSTEM**CREDIT	POLITICS AND GOVERNMENT**TREATIES

User Interest-English Books Parameters: n=10, m=5	
CPLEX12.2	**GREEDY SOLUTION**
English Language**Books and Literature	English Language**Books and Literature
Computers and the Internet**Books And Literature	Computers and the Internet**Books And Literature
Children and Youth**Books and Literature	Children and Youth**Books and Literature
Cooking and Cookbooks**Recipes	Cooking and Cookbooks**Recipes
Books and Literature**Book Trade	Books and Literature**Book Trade

User Interest-Computer Science Parameters n=10, m=5	
CPLEX12.2	**GREEDY SOLUTION**
Computers and the Internet**Mergers, Acquisitions and Divestitures	Computers and the Internet**Mergers, Acquisitions and Divestitures
Computers and the Internet**Books And Literature	Computers and the Internet**Books And Literature
Computer Security**Computers And the Internet	Computer Security**Computers And the Internet
Computers and the Internet**Stocks And Bonds	Computers and the Internet**Stocks And Bonds
Computers and the Internet**Computer Chips	Computers and the Internet**Children And Youth

User Interest-Stock Market Parameters n=10, m=5	
CPLEX12.2	**GREEDY SOLUTION**
Stocks and Bonds**Mutual Funds	Stocks and Bonds**Mutual Funds
Supermarkets**Food	Supermarkets**Food
Hedge Funds**Stocks and Bonds	Hedge Funds**Stocks and Bonds
Taxation**Stocks and Bonds	Securities and Commodities Violations**Stocks and Bonds
Research**Stocks and Bonds	Research**Stocks and Bonds

User Interest-Military Affairs Parameters n=10, m=5	
CPLEX 12.2	**GREEDY SOLUTION**
Airlines and Airplanes**Military Aircraft	Airlines and Airplanes**Military Aircraft
Two Thousand and One (Year)**United States Armament and Defense	Two Thousand and One (Year)**United States Armament and Defense
Politics and Government**Legislatures And Parliaments	Politics and Government**Legislatures And Parliaments
Terrorism**Two Thousand and One (Year)	Terrorism**Two Thousand and One (Year)
Terrorism**Accidents and Safety	Terrorism**Accidents and Safety

User Interest-Space Science Parameters n=10, m=5	
CPLEX 12.2	**GREEDY SOLUTION**
Space**Discovery (Space Shuttle)	Space**Discovery (Space Shuttle)
Space Shuttle**Discovery (Space Shuttle)	Space Shuttle**Discovery (Space Shuttle)
Accidents and Safety**Space Shuttle	Columbia (Space Shuttle)**Space Shuttle
Columbia (Space Shuttle)**Space	Columbia (Space Shuttle)**Space
Columbia (Space Shuttle)**Accidents And Safety	Columbia (Space Shuttle)**Accidents And Safety

Bibliography

[1] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *CIKM*, pages 316–321, 1999.

[2] Spärck Jones and Karen. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[3] Liliana Ardissono and Mark Maybury. Special issue on user modeling and personalization for tv. *Journal of User Modeling and User-Adapted Interaction*, 14, 2004.

[4] Ronald T. Fernndez and David E. Losada. Novelty detection using local context analysis. In *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277878.

[5] Filip Radlinski and Susan Dumais. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 691–692, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148320.

[6] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(6): 734–749, 2005.

[7] F. Qiu and J Cho. Automatic identification of user interest for personalized search. In *In Proceedings of the 15th International Conference on World Wide Web*, pages 727–736. ACM Press, New York, 2006.

[8] Boris Rousseau, Parisch Browne, Paul Malone, and Mchel Foghl. User profiling for content personalisation in information retrieval. In *19th Annual Conf. of ACM Symposium on Applied Computing (ACMSAC04)*, Nicosia, Cyprus, 2004. ACM.

[9] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[10] W. Bruce Croft and John Lafferty, editors. *Language Modeling for Information Retrieval*. Springer, New York, 2003.

[11] Iraklis A. Klampanos, Manning Christopher, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Information Retrieval*, 12(5):609–612, 2009.

[12] ChengXiang Zhai. Statistical models for information retrieval. In *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2008. doi: 10.2200/S00158ED1V01Y200811HLT001.

[13] Qiaozhu Mei, Hui Fang, and ChengXiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR*, pages 319–326, 2007.

[14] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. *Information Retrieval-Implementing and Evaluating Search Engines*. MIT Press, 2010. ISBN ISBN-13:978-0-262-02651-2.

[15] Didier Bourigault and Christian Jacquemin. Term extraction + term clustering: an integrated platform for computer-aided terminology. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 15–22, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/977035.977039. URL http://dx.doi.org/10.3115/977035.977039.

[16] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[17] Zhiguo Gong, Chan Wa Cheang, and Leong Hou U. Web query expansion by wordnet. In *In DEXA*, pages 166–175, 2005.

[18] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Query expansion using heterogeneous thesauri. *Inf. Process. Manage.*, 36(3):361–378, 2000.

[19] R. Attar and Aviezri S. Fraenkel. Local feedback in full-text retrieval systems. *J. ACM*, 24(3):397–417, 1977.

[20] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297, 1990.

[21] Shu-Cherng Fang and Sarat Puthenpura. *Linear optimization and extensions: theory and algorithms*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-915265-2.

[22] Erling D. Andersen. Linear optimization: Theory, methods, and extensions, 2001.

[23] IBM ILOG CPLEX Optimizer. Cplex12.2. URL http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[24] Guy E. Blelloch, Richard Peng, and Kanat Tangwongsan. Linear-work greedy parallel approximate set cover and variants. In *SPAA*. San Jose, California, USA, 2011.

[25] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Max-cover in map-reduce. In *WWW*, pages 231–240, 2010.

[26] Xiaoyong Liu and W. Bruce Croft. Cluster-based retrieval using language models. In *SIGIR '04 Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009026.

[27] Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8. doi: 10.1145/243199.243216.

[28] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[29] Kuo, C.-C., Glover F., and Dhir K. S. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences*, 24(6):1171–1185, 1993. doi: 10.1111/j.1540-5915.1993.tb00509.x.

[30] E. Erkut. The discrete p-dispersion problem, 1990.

[31] Ghosh J.B. Computational aspects of the maximum diversity problem, 1996.

[32] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, 1998. doi: 10.1145/290941.291025.

[33] Glover F., Kuo C.C., and Dhir K.S. Heuristic algorithms for the maximum diversity problem, 1998.

[34] Roberto Aringhieria, Maurizio Bruglieri, and Roberto Cordone. Semidefinite bounds for the maximum diversity problem, 2006.

[35] Glover F and Woolsey E. Converting the 0-1 polynomial programming problem to a 0-1 linear program, 1974.

[36] Roberto Aringhieri, Roberto Cordone, and Yari Melzani. Tabu search versus grasp for the maximum diversity problem. *4OR: A QUATERLY JOURNAL OF OPERATIONS RESEARCH*, 6:45–60, 2008. doi: 10.1007/s10288-007-0033-9.

[37] M.G.C. Resende, R. Mart, M. Gallego, and A. Duarte. Grasp and path relinking for the max-min diversity problem. *Computers & Operations Research*, 37(3):98 – 508, 2010. doi: 10.1016/j.cor.2008.05.011. URL http://www.sciencedirect.com/science/article/B6VC5-4SM62MG-1/2/860ea421487e930a56269486b33e02d8.

[38] Optiscom Project. Instances of maximum diversity, . URL http://heur.uv.es/optsicom/mdp/.

[39] Optiscom Project. Instances of maxmin diversity, . URL http://heur.uv.es/optsicom/mmdp/.

[40] Dice and Lee R. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.