Pattern-based Topic Models for Information Filtering

Yang Gao, Yue Xu, Yuefeng Li

Queensland University of Technology
Brisbane, Australia

Workshop SENTIRE, ICDM'2013

Contents

Background and Motivation

Information Filtering

Definition: Information Filtering (IF) is a system to remove irrelevant or unwanted information from an information or document stream based on document representations which represent users' interest.

Background and Motivation

Information Filtering

- Users' interest model (document model)
- Document relevance ranking

Traditional IF models:

- Term based models (BM25, SVM, Rocchio, etc.)
- Pattern based models (Closed pattern, sequential pattern)

Assumption: One user is only interested in a single topic

Statistical topic modelling:

Documents and collections can be represented by **multiple topics**, each topic can be represented by distributions of words.

Limitation: single words lack of distinguished semantic meaning to represent topic

Background and Motivation

Pattern-based Topic Model for IF

- Users' interest model (document model)
- Document relevance ranking

Assumption: One user is only interested in single topic

Limitation: single words lacks of distinguished semantic meaning to represent topic

Pattern-based Topic Model (PBTM)

Combines pattern mining with statistical topic modeling to generate more discriminative and semantic rich topic representations

- **Benefits of using pattern mining**
  - Discover hidden associations among words to represent the documents and the collection
  - Patterns carry more semantic meaning than single words.

Latent Dirichlet Allocation

- The statistical topic modeling technique has attracted great attention due to its robust and interpretable topic representations.
- The most popular used topic modeling method is LDA (Latent Dirichlet Allocation), and its various extensions.
  - Each document is a mixture of topics
  - Each topic is represented by distributions of words

Latent Dirichlet Allocation

topic models

Corpus Level

A collection of documents is represented by a number of topics. Each topic is represented by group of words with probabilities.

Document Level

Every document in this collection is represented by a distribution of the topics, each topic is represented by group of words with probabilities.

Word Level

In each document, every word is assigned with a topic and a probability. This word-topic assignment indicates which words are important to which topics at document level.

Latent Dirichlet Allocation – topic representation

Topic 0	Topic 10	Topic 11	Topic 12
method 0.043	data 0.437	method 0.072	classification 0.128
sample 0.042	mine 0.062	predict 0.064	feature 0.084
distribute 0.040	set 0.054	linear 0.028	accuracy 0.042
dimension 0.031	analyse 0.039	weight 0.025	class 0.036
parameter 0.028	application 0.033	kernel 0.025	method 0.036
estimate 0.027	method 0.030	variable 0.025	train 0.032
distance 0.025	real 0.023	model 0.025	performance 0.030
high 0.024	collection 0.014	dataset 0.025	selection 0.026
gene 0.023	recent 0.010	factor 0.025	machine 0.021
paper 0.023	synthetic 0.009	regression 0.023	state 0.020

- Common words in different topics produce ambiguous meaning across topics.
- Single words are not discriminative enough to represent the meaning of topics

PBTM for Topic Models

The process of proposed model:

Construct a transactional dataset from each topic's word based representation

Generate frequent patterns from the dataset to represent the topic: pattern based topic representation

TABLE I
EXAMPLE RESULTS OF LDA: WORD-TOPIC ASSIGNMENTS

Topic		Z_1		Z_2		Z_3
Document	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	w_1, w_9, w_8	0.2	w_7, w_{10}, w_{10}
d_2	0.2	w_2, w_4, w_4	0.5	w_7, w_8, w_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_1, w_7, w_5	0.3	w_7, w_3, w_3, w_2	0.4	w_4, w_7, w_{10}, w_{11}
d_4	0.3	w_2, w_7, w_6	0.4	w_9, w_8, w_1	0.3	w_1, w_{11}, w_{10}

PBTM for Topic Models

Example for topic Z_2

Construct transactional dataset

Generate pattern based topic representation

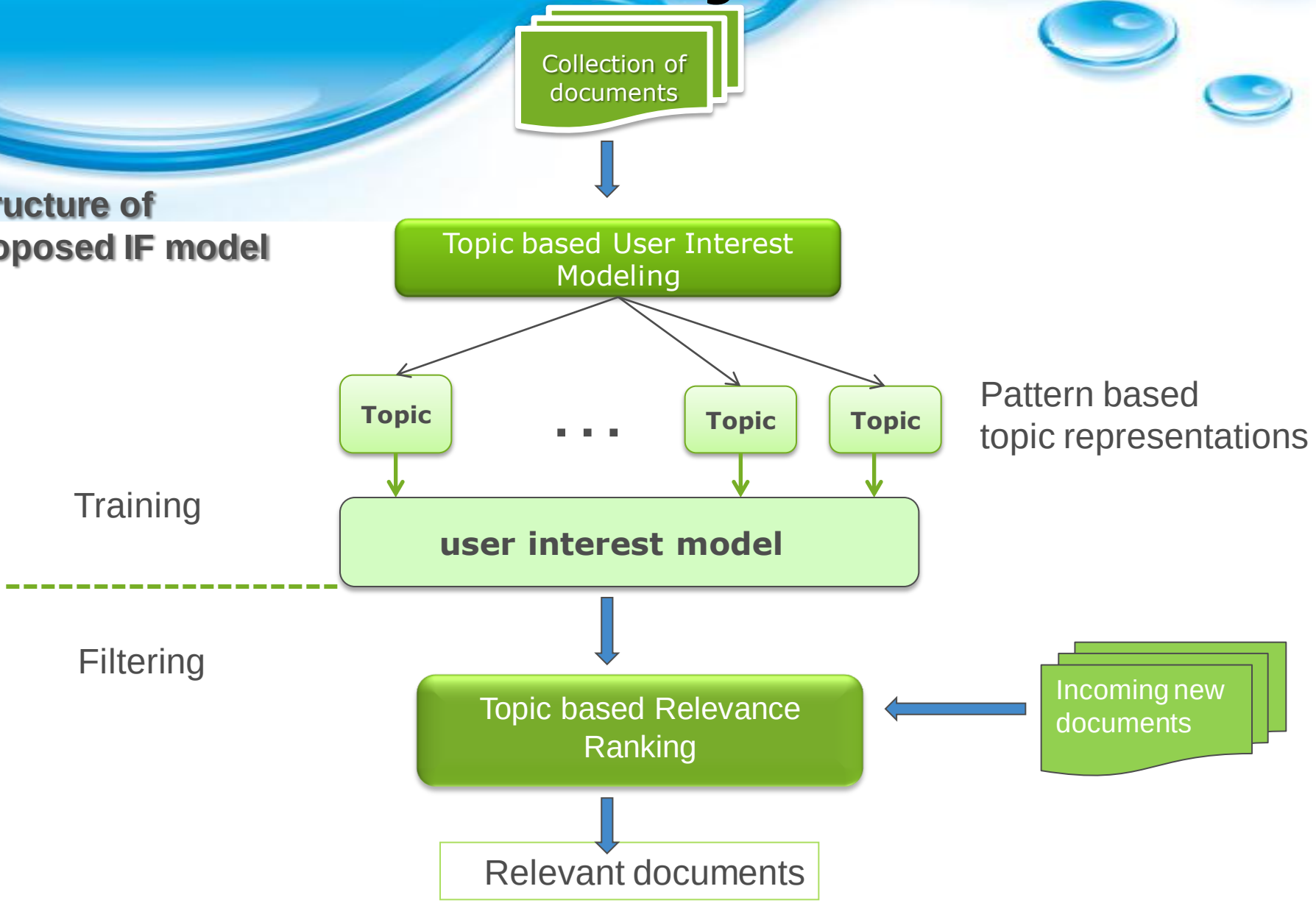
transaction	topic document transaction
1	$\{w_1, w_8, w_9\}$
2	$\{w_1, w_7, w_8\}$
3	$\{w_2, w_3, w_7\}$
4	$\{w_1, w_8, w_9\}$

Γ_2

Patterns	*supp*
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}\{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2

PBTM for Information Filtering

Structure of proposed IF model



PBTM for Information Filtering

Topic based User Interest Model

- For a document collection D and V pre-specified latent topics, from the results of LDA to D, generate V transactional datasets $\Gamma_1,\ldots,\Gamma_V$.

- Generate user interest model, $U = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$
  $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$ is a set of frequent patterns generated from Γ_i
  The patterns in $\mathbf{X}_{Z_i}$ represent what the user is interested in terms of topic Z_i

- θ_D represents topic distribution of D, which is used to represent the user's topic interest distribution. $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \cdots, \vartheta_{D,V}), \sum_{j=1}^{V} \vartheta_{D,j} = 1,$

PBTM for Information Filtering

Topic based Relevance Ranking

- **Pattern Specificity**:

  The specificity of a pattern X is defined as power function of the pattern length with the exponent less than 1, denoted as $spe(X)$, $spe(X)^m = a|X|^m$, a and m are constant real numbers. In this paper, $a = 1$, $m = 0.5$.

- **Topic Significance**:

  Let d be a document, Z_j be a topic in the user interest model, PA_{jk}^d be matched patterns, $k = 1,\ldots, n_j$, to document d, and $f_{j1}, \cdots, f_{jn_j}$ be the corresponding frequencies of the matched patterns within Z_j, the topic significance Z_j of to d is defined as:

$$sig(Z_j, d) = \sum_{k=1}^{n_j} spe(PA_{jk}^d) \times f_{jk} = \sum_{k=1}^{n_j} |PA_{jk}^d|^{0.5} \times f_{jk}$$

PBTM for Information Filtering

Topic based Relevance Ranking

- The new idea of the proposed model is to use multiple topics to represent a collection, and represent each topic using semantic patterns. We choose two widely used patterns, frequent patterns and closed patterns to represent topics. The two models are **PBTM_FCP** and **PBTM_FP**.

- **Document ranking**:

  For an incoming document d, the relevance of d to the user interest model is estimated by topic significance and topic distribution:

$$rank(d) = \sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,j}$$

PBTM for Information Filtering

Topic based Relevance Ranking

- **Document ranking**:

  For an incoming document d, the relevance of d to the user interest model is estimated by topic significance and topic distribution:

$$rank(d) = \sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,j}$$

$$rank(d) = \sum_{j=1}^{V} \sum_{k=1}^{m_j} | X_{jk}^{d} |^{0.5} \times \vartheta_{D,j}$$

Ranking in **PBTM_FP** model, X is the frequent pattern in user's interest model

$$rank(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} | c_{jk}^{d} |^{0.5} \times \vartheta_{D,j}$$

Ranking in **PBTM_FCP** model, c is the closed pattern in user's interest model

Evaluation

- **Hypothesis**:

  User interests involve multiple topics, document modelling by taking multiple topics into consideration can generate more accurate user information needs.

- **Data**:

  In Reuters Corpus Volume 1 (RCV1), there are total of 806,791 documents contained in 100 collections. We choose the first 50 collections in this paper.

- **Measures:**

  The effectiveness is assessed by five different measures: *top* 20, Mean Average Precision (MAP), break-even point (b/p), F_β (β=1) and 11 points.

Evaluation

- **Baseline Models:**

Pattern-based category

- FCP (frequent closed pattern model)
- PTM (a state-of-the-art sequential closed pattern model)
- n-Gram (n-gram phrases, n=3)

Topic-modelling-based category

- LDA (directly use topic distribution as user interest model and document representation)
- LDA_word (words associated with different topics are used to represent user interest needs and word frequency is used to represent topic relevance)

Evaluation

- **Results**

TABLE IV
COMPARISON OF ALL MODELS OVER ALL ASSESSING COLLECTIONS OF RCV1

Methods	*top20*	*b/p*	*MAP*	F_1
PBTM_FCP	**0.494**	**0.420**	0.424	**0.424**
PBTM_FP	0.47	0.402	**0.428**	0.424
LDA_word	0.447	0.410	0.415	0.423
LDA	0.337	0.295	0.308	0.339
change%	+5.1	+4.5	+3.1	+0.23
PTM	0.406	0.353	0.364	0.390
n-Gram	0.401	0.342	0.361	0.386
FCP	0.428	0.346	0.361	0.385
change%	+15.4	+19.0	+16.5	+8.7

Evaluation

- **Results**

TABLE IV
COMPARISON OF ALL MODELS OVER ALL ASSESSING COLLECTIONS OF RCV1

Methods	*top20*	*b/p*	*MAP*	F_1
PBTM_FCP	**0.494**	**0.420**	0.424	**0.424**
PBTM_FP	0.47	0.402	**0.428**	0.424
LDA_word	0.447	0.410	0.415	0.423
LDA	0.337	0.295	0.308	0.339
change%	+5.1	+4.5	+3.1	+0.23
PTM	0.406	0.353	0.364	0.390
n-Gram	0.401	0.342	0.361	0.386
FCP	0.428	0.346	0.361	0.385
change%	+15.4	+19.0	+16.5	+8.7

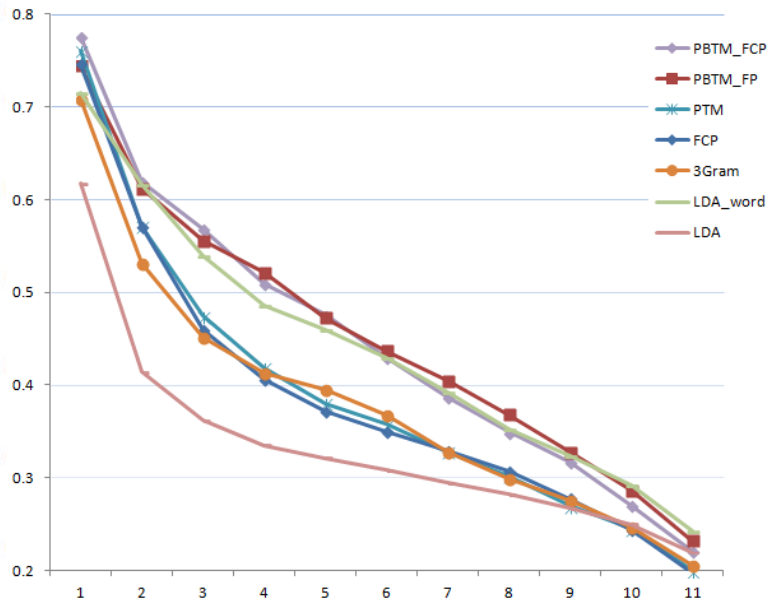


Fig. 2. Comparison between the proposed method and baseline models

Evaluation

- The experiment results clearly show that taking topics into consideration in generating user interest models can greatly improve the performance of information filtering.
- Pattern based topic models (PBTM) outperform word based topic model (LDA-word), which shows the benefit obtained by incorporating pattern mining into topic modeling, which is an important contribution of this paper.
- PBTM_FCP in most cases outperforms PBTM_FP, which indicates that using closed patterns to represent user interests is more accurate than using frequent patterns.
- The complexity of PBTM is determined by topic modeling or pattern mining, in most cases, by pattern mining. The complexity of efficient pattern mining methods such as FP-Tree has been proved acceptable in practice.

Summary and future work

- PBTM firstly generates pattern based topic representations to model user's information interests with multiple topics; then PBTM selects quality patterns for estimating the relevance of documents. The proposed approach incorporates the semantic topics from topic modeling and the specificity of the representative patterns. The proposed model has been evaluated by using RCV1 and TREC topics for the task of information filtering. Comparing with the state-of-the-art models, PBTM demonstrates excellent strength on document modelling and relevance ranking.

- In the future, we can select more discriminative and precise patterns for representing topics and document relevance.


Thank You !