News Selection with Topic Modeling

Cagri Toraman
Bilkent Information Retrieval Group
Computer Engineering Department
Bilkent University, Ankara, Turkey
ctoraman@cs.bilkent.edu.tr

There are numerous news articles coming to news aggregators and important news are selected to be presented on the front-page. There are two types of news selection for the front-page of news aggregators: personalized and public news recommendation (selection). This study examines public news recommendation that aims to satisfy all users' interest on the front-page. Public news recommendation is mainly done by meta-features like news popularity. A different approach that exploits the news content is introduced in this work. The main target is to select important (significant) news articles while providing diversification in the selected news topics. A new approach based on topic modeling is developed for this purpose. Results show that it is hard to achieve satisfactory level of precision when content-based public news recommendation is applied. However, precision of topic modeling-based approach is noticeably better than precision of random news recommendation. Topics of selected news are also diversified by using topic modeling.

News aggregators, news recommendation, news selection, topic modeling.

1. INTRODUCTION

There are numerous news articles coming to news aggregators from lots of news sources. Typical examples of news aggregators are Google News and Yahoo! News. Important news articles are selected (recommended) to present on the front-page (main-page) of news aggregators. The recommendation process should take scalability of system (millions of users) and amount of news churn (insertions and deletions) into account [1]. Therefore, there is a need for fast and effective news recommendation approaches.

What is important news and what makes news important? These questions are beyond the scope of this study. Instead, news aggregators simplify the task into two types of news recommendation on the front-page: personalized and public news recommendation.

Personalized news recommendation aims to satisfy a specific user's news interests while preserving importance of news. Mostly, a profile model associated with the user is generated and candidate news articles are filtered on user's profile model whenever user logs in to the system. User's past history and similar users' system activities are exploited to generate the model.

Public news aggregators simply assume that popular news articles are important. Popularity is mostly measured by meta-features like number of

clicks. This study examines public news recommendation, but with a different approach that exploits the news content instead of meta-features. The motivation of using news content is to explore rich text content to get useful information and to be adaptable for cases when there are no meta-features but only news content.

News aggregators should take the following factors into account while selecting (recommending) news for the front-page:

- *Importance ranking:* News should be ranked according to their importance. Importance is an abstract definition for public front-pages. Popularity is one measure to find importance. Importance in this study measures how a news article represents a cluster/topic.
- *Diversification:* News agenda should be presented in the front-page as much directions or viewpoints as possible. This study aims to increase topic variety on the front-page to obtain diversification.
- *Length of front-page:* News aggregators have a limited space to present the most important news articles while providing enough diversification on the front-page. This study examines 10-length front-pages.

This study develops a new method based on topic modeling that considers all of the above factors for public news recommendation. The following section

gives a brief related work on news recommendation. Section 3 explains the details of topic modeling-based approach. Section 4 gives the experiment setup and results. Section 5 concludes the study with some key future points.

2. RELATED WORK

Recommendation is mainly categorized into collaborative and content-based filtering. Collaborative filtering aims to exploit similar users' activity on the system whereas content of item (news article) is processed in content-based filtering. A typical collaborative filtering example for news recommendation is early version of Google News [Das et al., 2007]. Users with the similar click history are obtained and the news they read are recommended. Later, recommendation approach of Google News is changed to adapt both collaborative and content-based filtering [Liu et al., 2010]. A content-based approach that models user's information profile is mixed with the previous collaborative click history method. User profile is built on user's news interests by using news articles that particular user read before.

Content-based filtering of news articles is also examined in various studies. Common point of these studies is to create user profile implicitly or explicitly. Getting feedback from users after they read news articles is one way for explicit user profiling [Bilsus, Pazzani, 1999; Tan et al., 1998]. Other studies spy user activities to create user profile implicitly [Good et al., 1999].

This study examines content-based filtering in a different approach. Proposed method does not spy on user activities nor get feedbacks from user; thereby it is much like a selection method rather than a recommendation approach. The text content of news article is processed without any user information. It can be used for public front-pages.

Topic modeling has been recently studied widely and also applied to news domain [Newman et al., 2006]. The aim is to model a collection of texts into groups/topics. A popular topic modeling approach, Latent Dirichlet Allocation (LDA) [Blei et al., 2003], is used in this study. As far as I know, there is no study that adapts topic modeling into public news recommendation.

2. TOPIC MODELING-BASED NEWS SELECTION

Topic modeling-based news recommendation for public front-pages of news aggregators is illustrated in Figure 1. The following steps are the main tasks of the algorithm:

(i) Pre-processing news articles,

- (ii) Topic modeling of pre-processed news articles with LDA algorithm,
- (iii) Calculating document importance values using the topic model and ranking documents according to their importance,
- (iv) Calculating topic importance values using the topic model and ranking topics according to their importance,
- (v) Selecting (recommending) news articles using the rankings from (3) and (4) based on priority scheduling.

2.1 Pre-processing

Pre-processing includes stemming and removing stop-words. Stemming strategy is to use just first 5 characters of all words, which is approved to show good results in Turkish text [Toraman et al., 2011]. Most common word list in Turkish is obtained from [Can et al., 2008].

Figure 1: Main tasks of news selection with topic modeling.

2.2 Topic Modeling with LDA

Latent Dirichlet Allocation (LDA) algorithm assigns latent topics to each word in a given text collection representing the text collection in the vector space model [Salton et al., 1975]. Briefly, LDA determines two basic distributions that are word distribution for each topic and topic distribution for each document. These distributions can be used for determining topic label of each document and also ranking representative words for each topic. LDA algorithm learns a topic model; however this study does not use train-test paradigm, instead distributions obtained from the model are exploited to develop new importance models for documents and topics. A sample topic~word distribution obtained from LDA algorithm is listed in Table 1.

Table 1: A sample topic~word distribution obtained from LDA algorithm on news articles of 12.09.2009. Stemming first 5 characters is applied. English translations or explanations of Turkish words based on non-stemmed versions for each topic are given inside parentheses for giving inspiration on topic meanings.

Topic	First 5 representative words in topic-word distributions of each topic				
	#1	#2	#3	#4	#5
1	sel (flood)	sular (water)	kurta (save)	çalış (work)	yağış (rain)
2	bin (thousand)	dolar (dollar)	sanay (industry)	kredi (credit)	milyo (million)
3	iftar	numar (number)	araşt (research)	serib (seeding)	tenis (tennis)
4	şampi (champ.)	takım (team)	oyunc (player)	türki (turkey)	polon (poland)
5	polis (police)	kışı (people)	kızla (girl)	yaşın (aged)	genç (young)
6	fotoğ (photo)	beled (minis.)	sergi (exhib.)	türki (turkey)	paris
7	takım (team)	futbo (footb.)	oyunc (player)	maçın (match)	kazan (win)
8	insan (people)	kültü (culture)	zaman (time)	tarih (history)	mesel (issue)
9	türki (turkey)	çalış (work)	büyük (big)	getir (benefit)	yeni (new)
10	istan (istanbul)	bölge (zone)	felak (disaster)	yağış (rainfall)	vatan (country)
11	asker (soldier)	şehit (martyr)	terör (terror)	davut (Turkish foreign minister Davut-oglu)	ziyar (visit)
12	yargı (judg.)	cumhu (repub.)	demok (democ.)	mille (nation)	parti (party)
13	doğan (a Turkish media group)	vergi (tax)	hisse (share)	holdi (holding)	yayın (publish)
14	hasta (sick)	sağlı (health)	konya (a city in Turkey)	araşt (research)	hatal (mistake)
15	kadın (woman)	anlat (tell)	hatic (hatice)	çocuk (children)	hayat (life)

2.3 Calculating Document Importance

Assume number of documents in text collection N and importance value for each document d_i ($1 \leq i \leq N$) is calculated by using topic distribution value of each word in d_i . Topic distribution value of each word is obtained from the d_i 's topic distribution. This measures intuitively document's importance in its topic by calculating how words in d_i represent topic of d_i . Assume L_i as topic of d_i , $|d_i|$ as the number of words in d_i after pre-processing and t_{ij} as each word in d_i ($1 \leq j \leq |d_i|$). Distribution weight of word t_{ij} in topic L_i is w_{ij} and LDA gives this distribution as output. The following formula $imp(d_i)$ measures how important (representative) a document is:

$$imp(d_i) = \frac{\sum_{j=1}^{|d_i|} w_{ij}}{|d_i|} \quad (2.1)$$

Figure 2 displays small amount of words have larger distribution weights than most of words. The distribution in figure implies Power Law [Newman, 2005]. The formula (2.1) can then changed to calculate for only the k words that have high distribution weights as the following:

$$imp_k(d_i) = \frac{\sum_{j=1}^k w_{ij}}{|d_i|} \quad (2.2)$$

The value of k can be determined by Pareto Principle that implies 80-20 Law [Newman, 2005]. It means 80% of words that have high value of distribution weight are in the first 20% of all words distribution.

Figure 2: A sample topic-word distribution weights for an arbitrary topic obtained from a random day in dataset. Words are ranked in distribution weight order.

2.4 Calculating Topic Importance

Assume M as the number of topics. The value of M must be given to LDA algorithm as input. There are various studies to determine number of topics in a text collection. This study simply assumes M by inspring from clustering studies [van Rijsbergen, 1979] as the following:

$$M = \sqrt{(Total\ number\ of\ documents)/2} \quad (2.3)$$

For each topic L_i ($1 \leq i \leq M$), topic importance is calculated by using distribution weights of each word in topic distribution and also importance of documents in this topic. Note that all words appear in all topic distributions; but their weights would differ. Assume $|L_i|$ as total number of words in topic and t_{ij} as each word in topic ($1 \leq j \leq |L_i|$), D as total number of documents in topic L_i . Distribution weight of word t_{ij} in topic L_i is w_{ij} . The topic L_i 's importance $topic_imp_k(L_i)$ is calculated as follows:

$$topic_imp_k(L_i) = \frac{\sum_{j=1}^{|L_i|} w_{ij}}{|L_i|} + \frac{\sum_{j=1}^D imp_k(d_{ij})}{D} \quad (2.4)$$

Left side of the formula represents how much words in the topic are important by counting their distribution weights. Right side of the formula simply adds importance values of documents that are member of the topic.

2.5 Front-page News Selection Strategy

Using the topic and document importance values obtained from previous steps, priority scheduling-based news selection strategy is applied to satisfy both importance and diversification of news articles in a limited space of front-page.

Priority scheduling [Silberschatz et al., 2008] states CPU must serve waiting processes in a limited time. Each process have a priority and work time. CPU starts with the process that has the highest priority and serve until it finishes. The other processes are also served in the same manner. In this study, CPU is the proposed algorithm for public news recommendation and process is a topic. Each topic has a demand of working its most important news articles on the front-page. The proposed algorithm serves each topic's important news articles by considering topic's priority, topic's demand and the length of front-page (assumed 10 in this study).

The following steps describe the strategy for news selection to front-pages:

- (i) Ranked documents according to their importance are obtained as in (2.2).
- (ii) Each topic's priority is assigned as topic importance calculated as in (2.4).
- (iii) Each topic has a demand of presenting its important news articles. This demand value is calculated as the following:

- a. Calculate weight of importance (woi) for topic i as:

$$woi(L_i) = \frac{topic_imp(L_i)}{\sum_{j=1}^M topic_imp(L_j)} \quad (2.5)$$

- b. Calculate constant value that is needed to represent a news article in front-page:

$$c = \frac{1}{(length = 10)} = 0.1 \quad (2.6)$$

- c. Calculate demand of topic (dot) as:

$$dot(L_i) = \lfloor \frac{woi(L_i)}{c} + 0.5 \rfloor \quad (2.7)$$

- (iv) Priority scheduling puts the most important news article of the topic with the highest priority in the first place. Other slots are served by considering topic priorities and their demands.

A sample news selection based on this strategy is displayed in Figure 3. There are four different topics and their weight of importance values (2.5) and demand values (2.7) are given in the figure. Demand values (dot) are calculated as described in the strategy above. d_{ij} represents the j^{th} important news article in topic i .

3. EXPERIMENTS

3.1 Experiment Setup

LDA algorithm is implemented with MALLET library [MALLET, 2013]. Number of iterations in LDA are set to 1000 and Dirichlet parameters α and β to find distributions are set to their default values (0.01). Documents are represented in vector space model and algorithms proposed in this study are implemented in Java.

Dataset contains 15,844 Turkish news articles from 36 different days BilFront-2009 [BilFront, 2013]. Each day has approximately between 400 and 500 news articles except two days that have approximately between 200 and 300 news articles. Each news is labeled between 1 to 4 to show their

$woi(L_1) = 6.0$	$dot(L_1) = 3$
$woi(L_2) = 3.0$	$dot(L_2) = 2$
$woi(L_3) = 9.0$	$dot(L_3) = 5$
$woi(L_4) = 2.0$	$dot(L_4) = 1$
<i>Front-page representation (length = 10 news articles):</i>	
d_{31}	d_{32}
d_{33}	d_{34}
d_{35}	d_{11}
d_{12}	d_{13}
d_{21}	d_{22}

Figure 3: A sample news selection strategy based on priority-scheduling.

importance by three different annotators. Values 1 and 2 are for unimportant news whereas 3 and 4 imply important news.

For each 36 days in dataset, independent experiments are conducted three times and average is taken. This is because each day is assumed to have a separate news agenda. Random news selection is used as a baseline method and random experiments are repeated 20 times. We can not use click count information or some other meta-feature since dataset includes only the news content.

Evaluation of experiments are based on precision. The evaluation scores between 1 and 4 of annotators scaled to binary- important and unimportant by discreting scores equal and bigger than 2.5, and lower than 2.5, respectively. Then precision is calculated as follows:

$$precision = \frac{n_{\geq 2.5}}{l}$$

where $n_{\geq 2.5}$ is the number of selected news in front-page that has average annotator scores higher than 2.5 and l is the length of front-page. Precision values of 36 days are obtained and then dataset precision is calculated as follows:

$$dataset\ precision = \frac{\sum_{i=1}^{36} precision(g_i)}{36}$$

where $precision(g_i)$ is the precision of i^{th} day.

3.2 Experiment Results

Table 2 shows dataset precision results for different versions of proposed algorithm called NST (news selection with topic modeling).

Repeated words: Document importance calculations (2.1) and (2.2) are done by considering all words in the document regardless of their repetitions. Table 2 gives experiment results of (2.1) with repetitions (NST_{rep}) and without repetitions (NST). It seems repetitions are useful to consider in the calculations.

Most important k words: (2.2) considers only k words that have highest weights. Table 2 gives results of considering $k=500$ words ($NST_{2.2}$) and all words ($NST_{2.1}$). Using all words has worse result than using a part of words.

Pareto Principle: The value of k can be determined by Pareto Principle as the first 20% of all words. Table 2 lists experiments different k values and also using only Pareto value (NST_{pareto}). Results show that Pareto Principle seems to work well as a generalized decision.

News selection strategy: News selection can be done either as in Figure 3 ($NST_{priority}$) or by choosing the most important news article from each topic, which resembles democracy, called ($NST_{democracy}$). Table 2 shows that using priority scheduling based strategy works better than democracy.

Comparison with random selection: NST_{best} assumes repeated words, best k words with Pareto Principle, and priority scheduling-based strategy. Baseline method selects news articles randomly without any limitations except the length of front-page. It is seen that NST_{best} works much better than random selection in terms of precision.

4. DISCUSSION and CONCLUSION

This study introduces a novel algorithm called NST that provides topics modeling-based news selection for public front-pages of news aggregators. It does not recommend news articles by using user's or similar users' past history. Instead, NST just uses the news content.

Table 2: Experiment results for topic modeling-based news selection (NST) with different scenarios.

Decisions	Approach	Dataset Precision
Repeated words	NST_{rep}	0.25
	NST	0.24
Most important k words	$NST_{2.1}$	0.25
	$NST_{2.2}$	0.26
Pareto value	$NST_{k=10}$	0.23
	$NST_{k=100}$	0.25
	$NST_{k=250}$	0.25
	$NST_{k=500}$	0.26
	NST_{pareto}	0.25
Priority Strategy	$NST_{priority}$	0.30
	$NST_{democracy}$	0.25
Comparison with random	NST_{best}	0.31
	Random	0.15

Selecting (recommending) important news articles among a huge text collection is a hard task when only the news content is used. Experiment results show that proposed algorithm NST works better than random news selection. Proposed algorithm also provides diversified front-pages intuitively since it employs topic modeling and priority scheduling-based strategy selects news from various topics. Some future points are the followings:

- Topic and document importance can be calculated with different methods. For measuring document importance, traditional Information Retrieval approaches like *tf-idf* [van Rijsbergen, 1979] can be used.
- Different methods to find the number of topics should be examined. This study assumes a simple method inspiring from clustering studies.
- Other topic modeling methods like pLSA [Newman, Block, 2006] can be examined.
- Effects of different front-page lengths should be examined.
- Proposed algorithm can be compared with other content-based or collaborative filtering approaches.
- Diversity between topics is considered in this study. However, diversity between documents can be employed to improve the model.
- Novelty detection for news articles can be used to get better front-pages in terms of news agenda coverage.
- Experiments should be repeated for statistical tests to compare different configurations in terms of significance.

5. ACKNOWLEDGMENT

This study is supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) grant no. 111E030.

5. REFERENCES

- BilFront (2013) Bilkent Information Retrieval Group's News Portal Front-Page Selection Test Collection. Available from http://www.cs.bilkent.edu.tr/~canf/bilir_web/resources/Datasets.html (14.05.2013).
- Billsus, D., Pazzani, M. (1999) A hybrid user model for news story classification. In: *Proceedings of UM'99*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 99-108.
- Blei, D., Ng, A., Jordan, M. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022.
- Can, F., Koçberber, S., Balcik, E., Kaynak, C., Öcalan, H. C., Vursavas, O. M. (2008) Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*. 59(3): 407-421.
- Das, A. S., Datar, M., Garg, A., Rajaram, S. (2007) Google news personalization: scalable online collaborative filtering. In: *Proceedings of (WWW '07)*. ACM, New York, NY, USA, 271-280.
- Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J. (1999) Combining collaborative filtering with personal agents for better recommendations. In: *Proceedings of AAAI '99/IAAI '99*. Menlo Park, CA, USA, 439-446.
- Liu, J., Dolan, P., Pedersen, E. R. (2010) Personalized news recommendation based on click behavior. In: *Proceedings of IUI '10*. ACM, New York, NY, USA, 31-40.
- MALLET (2013) Machine learning for language toolkit. Available from <http://mallet.cs.umass.edu> (14.05.2013).
- Newman, M. E. J. (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46 (5): 323-351.
- Newman, D. J., Block, S. (2006) Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*. 57:753-767.
- Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M. (2006) Analyzing entities and topics in news articles using statistical topic models. In: *Proceedings of ISI'06*. Springer-Verlag, Berlin, Heidelberg, 93-104.
- Salton, G., Wong A., Yang C. S. (1975) A vector space model for automatic indexing. *Communications of the ACM* 18(11): 613-620.
- Silberschatz, A., Galvin, P. B., Gagne, G. (2008) *Operating System Concepts* (8th ed.). Wiley Publishing.
- Tan, A., Teo, C., Keng, H. M. (1998) Learning User Profiles for Personalized Information Dissemination. In: *Proceedings of IJCNN'98*. IEEE, 183-188.
- Toraman, C., Can, F., Kocberber, S. (2011) Developing a text categorization template for Turkish news portals. In: *Proceedings of INISTA 2011*, İstanbul, 379-383.
- van Rijsbergen, C. J. (1979) *Information Retrieval* (2nd ed.). Butterworths.