Article

JIS

Journal of Information Science
2015, Vol. 41(5) 676–685

DOI: 10.1177/0165551515589069
jis.sagepub.com
SAGE

A front-page news-selection algorithm based on topic modelling using raw text

Cagri Toraman
Computer Engineering Department, Bilkent University, Turkey

Fazli Can
Computer Engineering Department, Bilkent University, Turkey

Abstract

Front-page news selection is the task of finding important news articles in news aggregators. In this study, we examine news selection for public front pages using raw text, without any meta-attributes such as click counts. A novel algorithm is introduced by jointly considering the importance and diversity of selected news articles and the length of front pages. We estimate the importance of news, based on topic modelling, to provide the required diversity. Then we select important documents from important topics using a priority-based method that helps in fitting news content into the length of the front page. A user study is subsequently conducted to measure effectiveness and diversity, using our newly-generated annotation program. Annotation results show that up to seven of 10 news articles are important and up to nine of them are from different topics. Challenges in selecting public front-page news are addressed with an emphasis on future research.

Keywords

Diversity; document importance; front page; LDA; news selection; priority scheduling; topic importance; topic modelling

1. Introduction

The front page of a news aggregator, like Google News[1] or Yahoo! News,[2] is the showcase where readers expect to see significant news articles. With human-editor-based news aggregators, the burden of reading several news articles and selecting important ones is a challenging task. Editors may select worthless news unintentionally, or even according to their own points of view. As a result, intelligent algorithms that allow news aggregators to process news and select significant ones, need to be developed.

Given a news stream, M, that arrives periodically to a news aggregator, let s be the length of its front page (i.e. number of news articles in the front page); the problem is to select a set of important news articles $I \subseteq M$, $|I| = s$, which we call front-page news selection.

What interesting and important news is and what makes news interesting or important are questions that are beyond the scope of this study. The Community of Social Sciences tries to answer such questions. Eilders [1] states that readers favour articles with high values of news factors. News factors are characteristics such as unexpectedness, cultural proximity and reference to persons [2]. In this study, we simplify and generalize news factors as follows:

- *Importance ranking* – News should be ranked according to its importance. Importance is an abstract concept for public front pages. Popularity is one possible measure to quantify importance. Another interpretation of importance is to what degree a news article represents a cluster or topic.
- *Diversification* – The news agenda should be presented with as many viewpoints as possible. Viewpoints can be news categories (classes) or topics.

Corresponding author:
Fazli Can, Bilkent Information Retrieval Group, Computer Engineering Department, Bilkent University, Bilkent, Ankara 06800, Turkey.
canf@cs.bilkent.edu.tr

- *Length of the front page* – News aggregators have a limited space to present the most important news articles, while diversifying content as much as possible.

There are two types of news selection (news recommendation): personalized and public. Personalized news selection [3] aims at providing news according to the user's interest. A profile model associated with the user is typically generated and candidate news articles are filtered through the profile model whenever the user logs into the system. The 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. In this study, we examine news selection for public front pages. Popularity is mostly measured by meta-features, like number of clicks. Selecting important news using click counts is called click-based news selection. However, the number of clicks for a news article is counted during a long period of time, and is therefore not suitable for detecting breaking news. Moreover, the number of clicks cannot be quantified in environments that do not keep track of clicks. Thus, rather than meta-features, we focus on raw text (news content).

News articles in front pages can be diversified using their category or topic tags. However, our aim is not to use meta-attributes, but to leverage raw text for this purpose. Finding well-separated clusters, or topics, of news articles can be a solution while using only raw text. Topic modelling approaches [4] find separate clusters of documents for different topics and generate topic-word distributions. These words can be used to measure document and topic importance, while choosing different documents from varying topics to provide diversification. We present a novel approach that employs latent Dirichlet allocation (LDA) [5] to find diversified public front pages, while taking into consideration the importance of news within topics. LDA is a probabilistic topic-modelling algorithm that finds latent topics in a given text collection, and has been widely applied to various domains such as genetics and computer vision [4].

The contributions of this study are as follows. A novel algorithm to select public front-page news is introduced that considers the importance, diversity and length of the front page. We measure document importance and topic importance based on a statistical model that uses topic modelling to provide diversification. Then we select important documents from important topics using a priority-based method, to fit news content into the length of the front page. To the best of our knowledge, this is the first study that examines public front-page news selection using only raw text. We conduct a user study and measure the effectiveness and diversity of our algorithm, with our new annotation program. Annotation results show that up to seven of 10 news articles are tagged as important and up to nine of them are from different topics.

The rest of the paper is composed of the following sections. Related work on news selection, and the diversity of selections, are given in the next section. We present our algorithm in Section 3. In Section 4, we evaluate our algorithm based on a user study. We discuss challenges in selecting public front-page news in the same section. Section 5 concludes with some future research pointers in public front-page news selection.

2. Related work

2.1. News selection

To the best of our knowledge, news selection for public front pages, using only raw text, has not been studied before. However, there is a study that compares public front pages chosen by news editors and the interest of the social media crowd, but it does not present an algorithm for our task [6].

The task is similar to the traditional information-retrieval task, in which a set of documents and a query are given, and a subset of documents related to the query are returned by ranking according to their relevance to the query. The difference is that there is no query for selecting front-page news.

Selecting a subset from a document collection is a general task; we assume that news recommendation is the most related research area to front-page news selection. Recommendation systems are mainly divided into three categories [7].

In *content-based recommendation*, news content that is clicked or favoured by users is processed, and then news content that is similar to favoured ones is recommended. In its simplest form, similarity among news content is measured with metrics such as the cosine similarity [8]. There is also the content-based filtering approach, which creates a user profile implicitly or explicitly, and filters other news content according to the user profile. Getting feedback from news readers is one method of explicit user profiling [9]. Other studies track user activities to create a user profile implicitly [10].

Collaborative filtering aims to exploit similar users' activity on the system. A typical example of news recommendation using collaborative filtering is the early version of Google News [11]. Users with similar click history are fetched and news articles they read are recommended. Later, the recommendation approach of Google News changed to adapt both collaborative and content-based filtering [3]. The content-based approach that models users' information profiles is mixed with the previous collaborative, click-history method. The user profile is built on her news interests using news

articles that she read previously. This is an example of *hybrid recommendation* that aims to combine both advantages of content-based and collaborative method to provide more effective systems [12].

In terms of target community, recommendation systems are divided into two categories: *personal* and *public* recommendation. The former considers only a specific user while deciding on a recommendation. Collaborative filtering is a method of personal recommendation. The latter is harder to solve than personal recommendation, since there are many different user needs waiting to be satisfied. Content-based approaches can yield a solution for public recommendation.

The algorithm introduced in this study is an instance of content-based public recommendation. It does not spy on user activities, nor does it get feedback from users. Instead, the raw text of news articles is processed without any user information or meta-features, like click counts.

2.2. Diversity in document selection

There are algorithms for selecting a diversified set of documents among a given collection, based on measures such as maximal marginal relevance [13]. The application of such algorithms is examined in Clarke et al. [14]. Selection based on diversity is also studied in the context of publish/subscribe systems [15]. In such systems, documents are obtained in a given period of time and then a subset of them is selected by applying greedy search, to find the most diverse item. Recommendation systems that consider diversification [16] find documents of a similar interest to the user profile/query, while presenting diverse results from various topics.

In this study, we do not apply such approaches directly, since our algorithm handles different factors altogether, namely, importance, diversity and the length of the front page. For diversification, we utilize topic modelling that aims at finding latent groups/topics in a text collection. To the best of our knowledge, there is no study adapting topic modelling for public front-page news selection.

3. Front-page news selection based on topic modelling

An overview of our front-page news-selection approach is given in Figure 1. The main steps of our approach can be summarized as follows (each step is detailed in the following subsections):

- Given a news stream with M documents, *find topics*, with topic modelling, to provide diversity, and assign each document to a topic.
- *Find the importance of documents* and rank documents in each topic according to document importance.
- *Find the importance of topics*.
- *Select* the most important news articles, in the most important topics, with a priority-based method for fitting to the length of the front page (f).

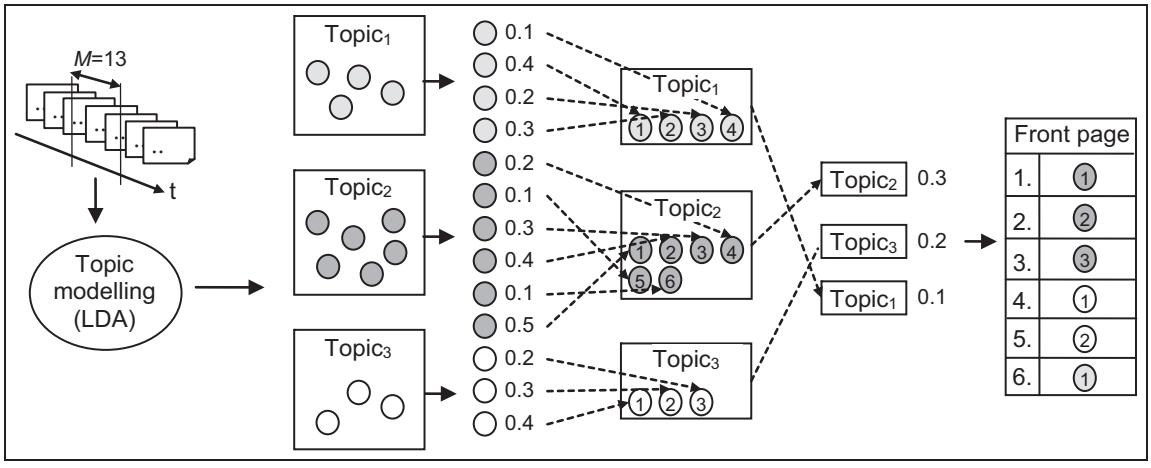

Figure 1. Overview of our front-page news selection approach. A news stream with M = 13 documents and a front page with a length (f) of 6 is given as an example. Documents are represented with circles and topics with rectangles. The numbers on the right of circles and rectangles are imaginary importance values for documents and topics, respectively. From left to right, dashed lines are used for document and topic ranking.

3.1. Finding topics

The LDA algorithm assigns latent topics to each word in a given text collection [5]. Briefly, LDA outputs topic-word (ϕ) and document-topic (θ) distributions. The word distribution for topic c (ϕ_c) estimates the probability of a word being generated by topic c. The topic distribution for document d (θ_d) estimates the probability of a topic being generated by document d. These are used for determining the topic of a document, and also the representative words for each topic. LDA learns a topic model, including these distributions, and can be used to predict the topic of a new document. However, in this study, we aim at utilizing distributions obtained from the model to estimate the importance of documents and topics.

3.2. Finding document importance

Let the number of documents in a given text collection be M and d_i's topic be c, and assume that the importance value for the document d_i ($1 \leq i \leq M$) is estimated by using ϕ_c of each word included in d_i. This measures, intuitively, d_i's importance by calculating how words in d_i represent d_i's topic. Let the topic assigned to d_i (i.e. highest probability in θ_i) be c, the number of words in d_i be T_i, t_{ij} be a word in d_i ($1 \leq j \leq T_i$), and the weight of word t_{ij} in ϕ_c be w_{ij}. The function *doc_imp(.)* measures how important (representative) a document is for its topic:

$$doc\_imp(d_i) = \left(\sum_{j=1}^{T_i} w_{ij}\right)/T_i \quad (1 \leq i \leq M) \tag{1}$$

We observe that a small number of words have high weights while others have low weights, which implies the power law [17]. This suggests that we trim low-weighted words, which are unimportant in the context of a given document and its topic, while calculating the *doc_imp(.)* function. We verify that ϕ follows a power-law distribution with the goodness-of-fit test [18]. The hypothesis of this test claims that ϕ is generated from a power-law distribution, and the test outputs a p-value that can be used for quantifying the validity of the hypothesis. If p-value is smaller than 0.1, then the hypothesis is rejected, which means that ϕ does not fit into a power-law distribution. If it is higher than 0.1, then ϕ is plausible for fitting a power-law distribution. For an arbitrarily chosen topic obtained from a random subset of a news collection that is used in evaluation, the p-value is obtained as 0.42. We observe similar patterns in other topics as well, and thus conclude that ϕ plausibly fits into a power law.

For trimming unimportant words, equation (1) is modified to quantify only for words in d_i that are seen in top k words of ϕ_c as follows:

$$doc\_imp(d_i) = \left(\sum_{j=1}^{T_i} \alpha.w_{ij}\right)/T_i,\ \alpha = \begin{cases} 1, & \text{if } t_{ij} \text{ is in top } k \text{ of } \phi_c \\ 0, & \text{otherwise} \end{cases} \quad (1 \leq i \leq M) \tag{2}$$

Note that the denominator is still document size, instead of the number of words seen in the top k of ϕ_c, which is to avoid overvaluing long documents with repeated high-weighted words (remember that T_i is the length of d_i). We determine the value of k as 20% of distribution length, which is obtained by the Pareto principle, which implies the 80–20 law [17]. This means that 80% of words that have a high weight are in the first 20% of ϕ_c.

3.3. Finding topic importance

Let S be the number of topics, which is given as an input to LDA. There are studies to determine the number of topics in a text collection [19], but for simplicity, assuming the number of documents in the given collection is M, we simply assume S's value is adapted from clustering studies [20] as $S = \sqrt{(M/2)}$ (rounded to the nearest integer).

For each topic c_i ($1 \leq i \leq S$), topic importance is calculated with the weights of words in ϕ_{c_i} and the importance values of documents that are assigned to c_i. Note that the same words appear in ϕ of all topics, but the weights of words in each ϕ are not necessarily the same. Let R be the total number of unique words in the given collection, r_{ij} be a unique word in topic c_i ($1 \leq j \leq R$), D_i be the total number of documents in topic c_i, d_{ij} be a document in topic c_i ($1 \leq j \leq D_i$) and the weight of r_{ij} in ϕ_c be w_{ij}. Assume that k is obtained by the Pareto principle as explained in the previous subsection. The importance of topic c_i is calculated with the *topic_imp(.)* function as follows:

$$topic\_imp(c_i) = \left(\sum_{j=1}^{R} \alpha.w_{ij}\right)/R + \left(\sum_{j=1}^{D_i} doc\_imp(d_{ij})\right)/D_i,\ \alpha = \begin{cases} 1, & \text{if } r_{ij} \text{ is in top } k \text{ of } \phi_{c_i} \\ 0, & \text{otherwise} \end{cases} \quad (1 \leq i \leq S) \tag{3}$$

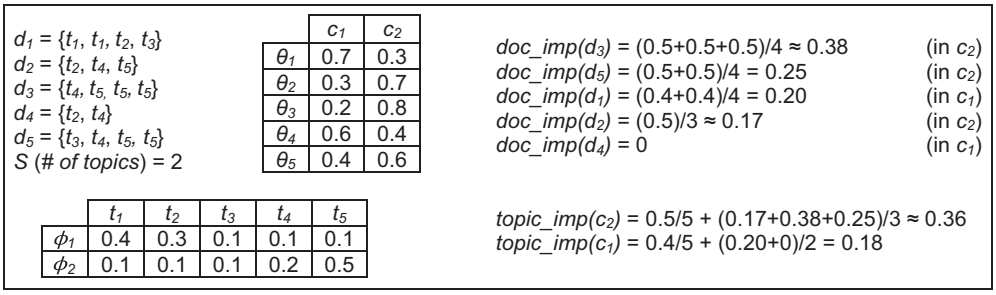

Figure 2. A sample document collection with $M = 5$ documents, $S = 2$ topics, and $R = 5$ unique words is given to demonstrate how to find document and topic importance. Recall that we use 20% of words in ϕ_c, then $k = 1$ for five unique words.

The first summation term in equation (3) estimates the importance of top k words of c_i by summing up their weights in ϕ_{c_i}. The second summation term estimates the importance of documents in c_i by summing up their *doc_imp(.)* values. Note that the number of words selected from R is exactly k, while the number of words selected from T_i in *doc_imp(.)* is equal to or lower than k, since a document may not necessarily include all top k words of its topic-word distribution. Figure 2 shows an example calculation to find document and topic importance by equations (2) and (3), respectively.

3.4. Priority-based news selection using document and topic importance

Having estimated topic and document importance values, we apply our priority-based method to provide diversification. Briefly, for a given front-page length, we select important news articles from various topics according to their priorities.

In operating systems, priority scheduling [21] solves the problem that the CPU must serve waiting processes in a limited time. Each process has a priority and time length. The CPU starts with the process with the highest priority and serves until it finishes. Other processes are then served in the same manner. In this study, we simulate that the CPU is our algorithm for public front-page news selection, and processes are topics. Each topic has a demand of placing its most important news articles on the front page. Our approach decides to serve important news articles in a topic by considering the topic's priority, demand and length of the front page.

Each topic has a priority value for being selected for the front page. Priority of c_i is calculated with the function *topic_pri(.)* as the portion of its importance value over all topic importance values where S is the number of topics:

$$topic\_pri(c_i) = topic\_imp(c_i) / \left(\sum_{j=1}^{S} topic\_imp(c_j)\right) \quad (1 \le i \le S) \tag{4}$$

Each topic also has a demand that shows how many important news articles this topic would like to place on the front page. The demand of c_i is calculated with the function *topic_dem(.)* – rounded to the nearest integer – where h is the constant to represent a news article's share in the front page and is calculated as $h = 1/f$ where f is the length of the front page:

$$topic\_dem(c_i) = topic\_pri(c_i)/h \quad (1 \le i \le S) \tag{5}$$

For the top place(s), our approach selects the most important news article(s) of the topic, which has (have) the highest priority, ordered by document importance. Other slots are served by topic priorities and their demands.

Assume that Figure 3 shows our front-page news selection strategy based on the same news collection given in Figure 2. Two front-page lengths are considered ($f = 3$ and $f = 4$). At the top of Figure 3, *topic_pri(.)* and *topic_dem(.)* values are calculated. For instance, $topic\_pri(c_1) = 0.33$ means that the first topic has a share of 33% on the front page and $topic\_dem(c_1) = 1$ means that the first topic demands 1 news article, according to its weight of importance. Since the highest topic importance value belongs to the second topic (inferred from the weight of importance values), we serve all of its demands in the first two and three slots of the front pages with $f = 3$ and $f = 4$, respectively. Then, the demand for the first topic is served in the remaining one slot. Note that each topic has a document ranking, according to document importance by equation (2), and ranking of the second topic is d_3, d_5 and d_2, while d_1 is the most important one in the first topic.

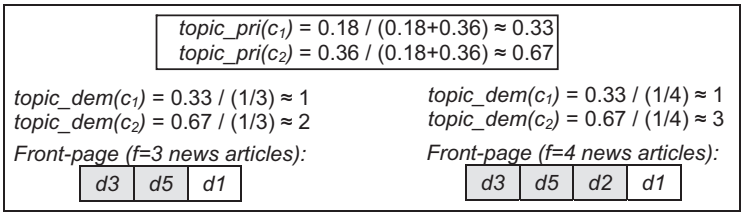

Figure 3. An illustration of selecting news for two different front-page lengths ($f = 3$ and $f = 4$), based on the same news collection given in Figure 2. Here, *topic_pri(.)* estimates the priority of a topic and *topic_dem(.)* finds how many documents a topic demands to place on front page. In the front-page representation, documents of the second topic are shaded in grey.

Algorithm 1. Pseudocode for our front-page news-selection approach.
Input: News collection with M documents, front-page length f
Output: Front-page F[..] that includes f news articles in ranked order.

```
    // Find topics (see Section 3.1)
1   Number of topics, S ← √(M/2)
2   Get topic-word distribution (φ) and document-topic distribution (θ) by topic modelling
    // Find importance of documents (see Section 3.2)
3   For i = 1…M do
4     Assign document d_i to the topic with the highest weight in θ_di
5     Get importance of d_i // (see Eq. 2)
6   End-for
    // Find importance of topics (see Section 3.3)
7   For i = 1…S do
8     Get importance of topic c_i // (see Eq. 3)
9     Rank documents in c_i according to their importance
10  End-for
    // Find priority and demand of topics (see Section 3.4)
11  For i = 1…S do
12    Get priority and demand of c_i // (see Eq. 4 and 5)
13  End-for
    // Select f news articles for front-page (see Section 3.4)
    // Consider topics in a ranked order according to their priority
14  While F has empty slot do
15    Consider the next topic
16    k ← demand of this topic // (see Eq. 5)
17    Place top k news articles from this topic to F
      // Note that available empty slots in F can be less than k
      // For example, if there are two slots available and k = 3, then select the top 2
18  End-while
```

Our front-page news-selection approach is given in Algorithm 1.

4. Evaluation

To the best of our knowledge, there is no gold-standard dataset of news articles, with labels of importance; therefore, we conducted a user study to evaluate the effectiveness of our algorithm in terms of document importance and topic diversity. Since we are not aware of any baseline algorithm for public front-page news selection with which we can compare our algorithm, the user study evaluates only our front-page news-selection algorithm. However, we have an additional user study to find the effectiveness of random news selection, and compare it with our algorithm in discussion. In this section, we explain the details and results of our user study using our newly-generated annotation program.

4.1. Setup

Approaches used to select front-page news are evaluated over a dataset that includes labels of whether a document is important or not; therefore, traditional metrics such as precision and recall can be measured. However, we are not aware of any labelled dataset suitable for our task. Yahoo! published a test collection including number of clicks for news articles in Yahoo! Front-page Today Module [22]; however this dataset does not include news article content and thus,

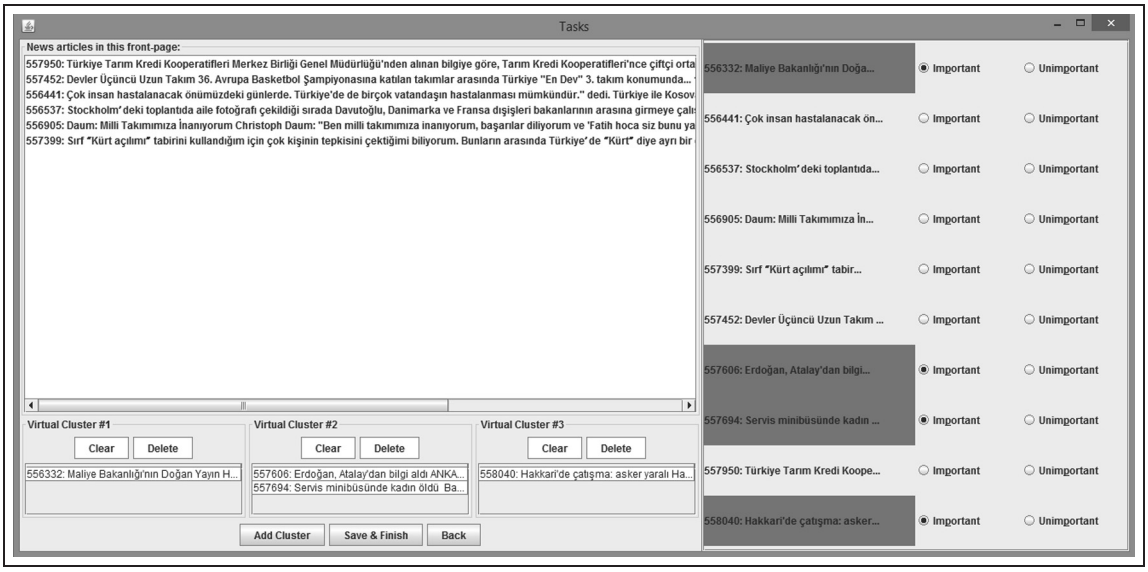

Figure 4. A sample screenshot from the annotation program. The front page has 10 news articles, and initially all are listed in the top-left panel. Annotators are asked to (a) assess the importance of each news article in the right panel and then (b) drag and drop each news article into a virtual cluster (topic) in the bottom-left panel. In this figure, all news articles are labelled as important or unimportant. However, diversity annotation is not yet completed. Four news articles (shaded in grey in the right panel) are dropped into three virtual clusters, and six news articles are still waiting for drag–dropping.

cannot be used in our study and similar studies that would examine methods using news content. Instead we use a non-labelled news collection, including 15,844 news articles that were obtained from the *Milliyet*[3] newspaper on 36 different days between 9 September 2009 and 31 October 2009, in which we obtain the cleanest raw text, via the Bilkent News Portal.[4] Since news articles were obtained from a real-time RSS resource, the number of news articles for each day differs.

News articles are first pre-processed by stemming and removing stopwords that are common words in Turkish. The stemming strategy is to use just the first five characters of all words, which is shown to yield good results in Turkish text [23]. The stopword list used in this study is a slightly extended version of the one obtained from Can et al. [24]. The LDA algorithm is implemented with MALLET library [25].

Since there are 36 days in the news collection, we run our algorithm for each day and get 36 different front pages. The front-page length is set to 10. We then conduct a user study to capture the importance and diversity of news articles selected by our algorithm.

4.2. The user study

We create an annotation program, with a user-friendly interface, capable of providing the label importance, while maintaining the diversity of the front page. A total of 19 graduate and undergraduate students are selected to be volunteer annotators. Before starting the annotation process, all annotators read the user manual that explains background information about the domain, tasks they will encounter, and how to use the annotation program.

Figure 4 shows a sample screenshot from the annotation program. In the top-left panel, news articles are listed by their snippets of 200 characters. Full text is accessible upon double-clicking on snippets. Annotators have two tasks that must be accomplished consecutively:

(1) First, annotators have a binary decision; whether given news articles are important or not, based on their own interests in the right panel.
(2) After deciding for all news articles in the right panel, they determine which news article belongs to which topic, that is, the diversity of a front page, using the bottom-left panel. However, there are no pre-defined topics in this task. Instead, they drag and drop each news article into a virtual cluster whose aim is to collect news articles belonging to the same topic. We call them virtual clusters since they have no descriptive title. Annotators add new, delete existing or change contents of virtual clusters if needed.

Table 1. Average, median, standard deviation, minimum and maximum of *ann_imp* and *ann_div* scores are listed for the user study of 19 annotators.

Type	Average	Median	Standard deviation	Minimum	Maximum
ann_imp (annotator importance)	0.52	0.51	0.10	0.27	0.70
ann_div (annotator diversity)	0.76	0.83	0.15	0.51	0.94

In the annotation screen of Figure 4, the importance annotation is completed since all news articles are labelled as important/unimportant in the right panel; however, diversity annotation is not finished. Four news articles are put into three virtual clusters (shaded in dark grey in the right panel), and six news articles are still waiting for drag–dropping.

In the user study, each annotator is assigned to the same tasks. Annotators have to complete 36 different front-page annotation processes. They may quit the program during a task and continue from the same state of annotation whenever they would like to do so. Also they can redo annotations that they finished earlier.

After all annotations are done, the importance and diversity of a front page are measured by *ann_imp(.)* and *ann_div(.)*, respectively. Let p_d be the total number of positive decisions (i.e. important news articles) for the annotator a, where d is a day in the given news collection. The importance of a front page of d, with length s, is $imp_a(d) = p_d/s$. Then, the importance of all front pages annotated by a is measured by the function *ann_imp(.)*:

$$ann\_imp(a) = (\sum_{i=1}^{36} imp_a(d_i))/36 \tag{6}$$

Similarly, let c_d be the total number of virtual clusters for day d. For the annotator a, diversity of a front page of d, with length s, is $div_a(d) = c_d/s$. Then, the diversity of all front pages annotated by a is measured by the function *ann_div(.)*:

$$ann\_div(a) = (\sum_{i=1}^{36} div_a(d_i))/36 \tag{7}$$

Both *ann_imp(.)* and *ann_div(.)* have a value between 0 and 1. The higher the importance and diversity that annotators achieve, the more important and diverse front pages our algorithm is meant to find.

Since there are more than two annotators, we calculate Fleiss's kappa [26] to estimate consistency between annotators. For importance and diversity annotations, Fleiss's kappa is calculated as 0.29 and 0.09, respectively. According to the interpretation given by Landis and Koch [27], there is a fair agreement for annotations of importance while annotations of diversity have slight agreement. Not having a strong agreement can be attributed to differences among user interests, which is an expected observation in public front-page news selection.

4.3. Results and discussion

Results of the user study are summarized in Table 1 that lists the average, median, standard deviation, minimum and maximum of annotator importance and diversity when 19 annotators are considered. For the sake of simplicity, details of annotation results are not given in this paper, but are available online.[5]

In the best case, we provide front pages including up to 70% important news articles, while up to 94% of them belong to different news topics. In the average case, our approach finds front pages including 52% important news articles while 76% of them belong to different news topics.

Note that front-page length is 10 in the user study. Front pages with a large number of news articles are difficult to annotate for importance and diversity; this is because annotations require the content of all news articles to be read. The more news articles are involved, the more information should be remembered.

Our results are not compared with any of those of other approaches to select public front-page news, since we are not aware of any similar study or method for our task. Click-based news selection is a possible solution, but no benchmarking is possible, since to the best of our knowledge there is no dataset that includes both click counts and raw text. However, we can compare the results of our approach with random news selection. For this purpose, in an additional user study, four annotators other than the previous ones are asked to assess the importance of all 15,844 news articles and only 1315 (~8%) of them were assessed as important. Details of additional-annotation results are available online.[6] Thus, if random

news selection is used, it is expected that approximately 8% of news articles on a front page will be important. This shows that our 52% success rate is 6.5 times more effective than that of random news selection.

One may also think of other solutions for this task, such as applying machine learning. In machine learning, a training model is learned by a classification algorithm, and then documents that have no class information are labelled by the learned model. Such an approach has obstacles. First, if machine learning is used for selecting public front-page news, a test collection, including gold standard is initially needed. However, there is no such test collection for such a task. Moreover, since the number of important news articles is much less than that of unimportant ones, there would be a bias towards unimportant news. Lastly, the news agenda would change in a news portal as new documents arrive. Since previously learned models show decreased performance on recent news agenda, a novel model should be learned based on previous ones. However, only the initial training model is learned with a gold standard. Performance would gradually decrease for recently learned models, based on previous news.

Another possible solution can be click-based news selection. Using this approach, one can measure the popularity of news articles easily by quantifying the number of clicks. However, initially the number of clicks does not exist, and also, misleading click information might be generated by robots, click spam, etc.

5. Conclusion

This study examines public front-page news selection by means of raw text, or news content, only. We develop a novel approach that finds topics with topic modelling to provide diversity, and then ranks documents according to their importance. Our algorithm selects the most important news articles in the most important topics with a priority-based method for fitting to the length of the front page. LDA is employed for topic modelling; however, other topic modelling methods can also be used [4]. We develop an annotation program for the purpose of conducting a user study. The study, employing 19 annotators, is conducted using 15,844 news articles to evaluate our method's performance. It shows that our topic modelling-based approach for public front-page news selection encourages the use of only raw text. In the best case, 70% of news articles are important and 94% are of different topics. Moreover, in the average case, 52% of news articles are important; this is about 6.5 times more effective than the 8% success rate of random news selection. Also, 76% of news articles are of different topics, in the average case.

In this study, importance and diversity are quantified by topic modelling; however, other methods can also be used. Topic tracking and novelty detection can be adapted for improving diversity and likewise, named-entity recognition can be used for improving news selection. In future studies, there is a need for labelled datasets, including both news content and number of clicks.

After providing a proper stopword list and stemmer, our method is language- and domain-independent and is suitable for similar, text-based applications such as blog, review and intelligence-report aggregators.

Acknowledgements

We thank our colleagues, friends and students for their annotations. We appreciate the comments of Volkan Yazar for the presentation of the manuscript.

Funding

This work is partly supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under grant number 111E030. Any opinions, findings and conclusions or recommendations expressed in this article belong to the authors and do not necessarily reflect those of the sponsor; therefore, no official endorsement should be inferred.

Notes

1. http:// news.google.com
2. http:// news.yahoo.com
3. http://www.milliyet.com.tr
4. http://139.179.21.201/PortalTest
5. http://cs.bilkent.edu.tr/~ctoraman/frontpage/annotation_results.pdf
6. http://cs.bilkent.edu.tr/~ctoraman/frontpage/additional_annotation.pdf

References

[1] Eilders C. The role of news factors in media use. Discussion Papers, Research Unit, The Public and the Social Movement. Social Science Research Center Berlin (WZB), 1996, pp. 96–104.

[2] Galtung J and Ruge MH. The structure of foreign news. *Journal of Peace Research* 1965; 2: 64–91.

[3] Liu J, Dolan P and Pedersen ER. Personalized news recommendation based on click behavior. In: *Proceedings of the 15th ınternational conference on ıntelligent user ınterfaces*, Hong Kong. New York: ACM, 2010, pp. 31–40.

[4] Blei DM. Probabilistic topic models. *Communications of the ACM* 2012; 55: 77–84.

[5] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003; 3: 993–1022.

[6] Zubiaga A. Newspaper editors vs the crowd: On the appropriateness of front-page news selection. In: *WWW companion volume*. International World Wide Web Conferences Steering Committee/ACM, 2013, pp. 879–880.

[7] Herlocker JL, Konstan JA, Terveen LG and Riedl JT. Evaluating collaborative filtering recommender systems. *ACM Transactions of Information Systems* 2004; 22: 5–53.

[8] Ahn J-w, Brusilovsky P, Grady J, He D and Syn SY. Open user profiles for adaptive news systems: help or harm? In: *Proceedings of the 16th ınternational conference on World Wide Web*, Banff, Canada. New York: ACM, 2007, pp. 11–20.

[9] Billsus D and Pazzani MJ. A hybrid user model for news story classification. In: *Proceedings of the seventh ınternational conference on user modeling*, Banff, Canada. New York: Springer, 1999, pp. 99–108.

[10] Good N, Schafer JB, Konstan JA et al. Combining collaborative filtering with personal agents for better recommendations. In: *Proceedings of the sixteenth national conference on artificial ıntelligence and the eleventh ınnovative applications of artificial ıntelligence conference ınnovative applications of artificial ıntelligence*, Orlando, FL. American Association for Artificial Intelligence, 1999, pp. 439–446.

[11] Das AS, Datar M, Garg A and Rajaram S. Google news personalization: scalable online collaborative filtering. In: *Proceedings of the 16th international conference on World Wide Web*, Banff, Canada. New York: ACM, 2007, pp. 271–280.

[12] Chu W and Park S-T. Personalized recommendation on dynamic content using predictive bilinear models. In: *Proceedings of the 18th ınternational conference on World Wide Web*, Madrid. New York: ACM, 2009, pp. 691–700.

[13] Carbonell J and Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st annual ınternational ACM SIGIR conference on research and development in ınformation retrieval*, Melbourne. New York: ACM, 1998, pp. 335–336.

[14] Clarke CLA, Kolla M, Cormack GV et al. Novelty and diversity in information retrieval evaluation. In: *Proceedings of the 31st annual ınternational ACM SIGIR conference on research and development in ınformation retrieval*, Singapore. New York: ACM, 2008, pp. 659–666.

[15] Drosou M and Pitoura E. Diversity over continuous data. *IEEE Data(base) Engineering Bulletin – DEBU* 2009; 32: 49–56.

[16] Ziegler C-N, McNee SM, Konstan JA and Lausen G. Improving recommendation lists through topic diversification. In: *Proceedings of the 14th ınternational conference on World Wide Web*, Chiba, Japan. New York: ACM, 2005, pp. 22–32.

[17] Newman M. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 2005; 46: 323–351.

[18] Clauset A, Shalizi CR and Newman MEJ. Power-law distributions in empirical data. *SIAM Reviews* 2009; 51: 661–703.

[19] Teh YW, Jordan MI, Beal MJ and Blei DM. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 2006; 101: 1566–1581.

[20] Mardia K, Kent J and Bibby J. *Multivariate Analysis,* 1st edn. London: Academic Press, 1979.

[21] Silberschatz A, Galvin PB and Gagne G. *Operating System Concepts*. Chichester: Wiley, 2008, p. 992.

[22] Yahoo! Research. Yahoo! Labs Datasets – Front-page Today Module Click Dataset, http://webscope.sandbox.yahoo.com/catalog.php?datatype=r (2015, accessed 5 May 2015).

[23] Toraman C, Can F and Kocberber S. Developing a text categorization template for Turkish news portals. In: *2011 International symposium on ınnovations in ıntelligent systems and applications (INISTA)*, 2011, pp. 379–383.

[24] Can F, Kocberber S, Balcik E, Kaynak C, Ocalan HC and Vursavas OM. Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology* 2008; 59: 407–421.

[25] McCallum AK. MALLET: Machine learning for language toolkit, http://mallet.cs.umass.edu (2015, accessed 5 May 2015).

[26] Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; 76: 378–382.

[27] Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.