# Temporal Corpus Summarization Using Submodular Word Coverage

Ruben Sipos
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
rs@cs.cornell.edu

Adith Swaminathan
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
adith@cs.cornell.edu

Pannaga Shivaswamy
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
pannaga@cs.cornell.edu

Thorsten Joachims
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
tj@cs.cornell.edu

## ABSTRACT

In many areas of life, we now have almost complete electronic archives reaching back for well over two decades. This includes, for example, the body of research papers in computer science, all news articles written in the US, and most people's personal email. However, we have only rather limited methods for analyzing and understanding these collections. While keyword-based retrieval systems allow efficient access to individual documents in archives, we still lack methods for understanding a corpus as a whole. In this paper, we explore methods that provide a temporal summary of such corpora in terms of landmark documents, authors, and topics. In particular, we explicitly model the temporal nature of influence between documents and re-interpret summarization as a coverage problem over words anchored in time. The resulting models provide monotone sub-modular objectives for computing informative and non-redundant summaries over time, which can be efficiently optimized with greedy algorithms. Our empirical study shows the effectiveness of our approach over several baselines.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

## General Terms

Design, Theory

## Keywords

summarization, temporal, submodular


*CIKM'12,* October 29–November 2, 2012, Maui, HI, USA.


## 1. INTRODUCTION

From news and blogs to twitter feeds, and from research papers to patents, we are accumulating unprecedented amounts of text in digital form. Advances in storage technology have allowed us to maintain complete records of these text streams, and information retrieval research has developed excellent tools for accessing individual documents in the resulting collections. However, our ability to analyze and interpret archives on a macroscopic level is still limited. Macroscopic questions one may ask about a collection range from the creation of a timeline of influential documents or authors, to the automatic summarization of the main chains of discussion.

To answer such macroscopic questions about a corpus of text documents, we draw upon methods from document summarization (see [14]). Instead of summarizing a single (or small number of) individual documents using extracted sentences, we aim to summarize a collection using extracted documents, authors, or keywords. This shift implies substantial differences in what constitutes a meaningful summary. In particular, time is more important for the creation of corpus summaries than it is for conventional summaries, and we argue that corpus summaries should reflect the influence that a document or author had on the future development of the collection. Therefore, our summaries take the form of timelines, where components of a summary are defined with respect to intervals or points of time.

More specifically, we formulate several variants of the corpus summarization problem. First, we seek to identify $k$ documents that had the largest influence on the content of the corpus. Second, for each point in time, we seek to identify those documents that were most influential for that time. Third, we similarly identify the most influential authors for each time-point. And fourth, we identify key phrases at each time-point that were influential and represent a coherent segment of the corpus.

All four corpus-summarization problems will be formulated as coverage problems, where we approximate coverage of abstract concepts through coverage of words in time. For conventional summarization and diversified retrieval, coverage approaches [21, 28, 16, 23, 19] and, more generally, submodular summarization methods [10] represent the state of the art. In particular, they provide an elegant model of the relevance/redundancy trade-off inherent in all summarization problems. The key technical challenge for the problem of corpus summarization is the ability to model time-points and time-intervals effectively, without sacrificing the computational approximation guarantees that the greedy algorithm provides for sub-

modular function maximization [13, 8]. Our approach models the fact that different ideas have varied novelty, and that the influence of an idea changes over time. An important feature of our approach is that it does not rely on observed or inferred link structure between documents, but requires only the time-stamped document text. This makes our approach applicable to a wide range of corpora for which citation information is not available or not reliable. Furthermore, it allows us to use citation information as "ground truth" for quantitative evaluation. On three scientific corpora, we perform such evaluations, compare to several baselines and find that our methods provide qualitatively interesting results.

## 2. RELATED WORK

Our approach in this paper is based on the idea of extractive summarization; in this approach, a summary is created by selecting smaller units (i.e. sentences) [3, 23, 10]. Corpus summarization has similarities to extractive document summarization: instead of a document to summarize, we have a corpus and instead of selecting sentences, we select entire documents. Extractive summaries can be obtained by maximizing the coverage of words over the corpus. This idea has been used both in document summarization [23] and information retrieval [28]. While existing approaches focused on covering words as a proxy for concepts, we extend the notion to identify documents from a corpus that had substantial influence on the future development of the corpus. This gives us a more versatile framework than, say, selecting sentences based on time-dependent event weights [25].

There are several approaches that deal with the corpus as a whole in an attempt to improve its accessibility. Recommender systems strive to suggest what to read next based on past behavior and personal preferences [7] or on a query set consisting of example papers [6]. A related approach uses collaborative filtering and/or content filtering [24]. The goal of recommender systems, broadly, is to suggest documents based on a query; the goal of our work is to understand how a corpus evolved over time and which authors, papers and key-words demonstrated their influence during the evolution.

There has been previous work on modelling and visualizing text corpora. On the macro level, we can describe a corpus in terms of topics [2]. There has also been work [4] which show trends in topics, indicating the main turning points. A different approach considers intra-corpus relations and describes influence between documents [18]. The disadvantage of these approaches is that the units being summarized (e.g. documents) and being visualized (e.g. turning points) are distinct objects. Bridging this disconnect is an important aim for summarization systems since providing an explicit guide in the visualization (e.g. by providing the influential documents in the corpus) helps a user become familiar with the novel content in a corpus.

Temporal text mining [20] is another popular area of research. There has been work [1] which views a news topic as a sequence of events and selects those events that are relevant and novel to form the summary. Further, timelines of events can be created [27, 22] which show the major developments in a news topic. There is also work on event threading [12] where events are not viewed as a flat hierarchy of topics; these approaches model the dependencies between events. In addition, temporal features can also be used when doing document summarization [9] to improve the performance. Our work implicitly models sequences of events over time, but always presents explicit examples that highlight influential contributions. The main insight of this work is that explicitly modelling the temporal context in which words appear in documents provides a simple and very effective approximation to the flow of ideas in a corpus.

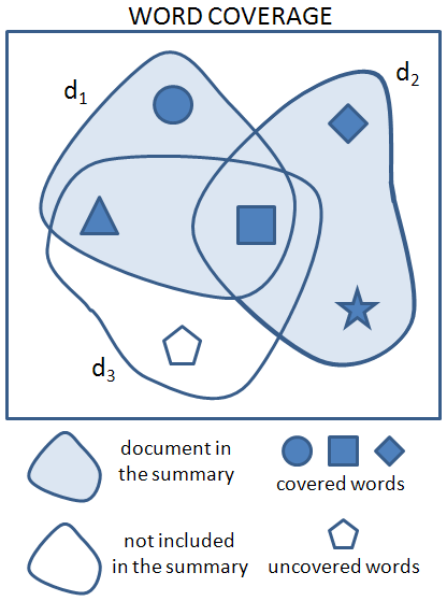


**Figure 1: Illustration of word-coverage objective.**

## 3. SUMMARIZATION AS COVERAGE

In this paper, we explore several variants of the corpus summarization problem, providing a temporal summary of the corpus in terms of landmark papers, authors and key-phrases. All approaches are formulated as maximum coverage problems, which have been found to provide elegant and effective methods for conventional summarization and diversified retrieval problems [10, 28, 21]. We start by reviewing the coverage-based summarization idea in the remainder of this section, and then extend it to corpus summarization in Section 4.

### 3.1 Information Coverage as Word Coverage

Coverage-based summarization methods make a direct analogy between a summary covering the information content in the object to be summarized, and maximum coverage problems as defined in theoretical computer science [8]. The key assumption of coverage-based summarization methods is that coverage of words can be used as a proxy for the coverage of information content. By achieving a good coverage over words, the word coverage approach aims to select a summary which covers different topics. In this way, coverage-based summarization methods elegantly avoid redundancy and promote diversity.

While document summarization involves selecting a diverse set of sentences from it, the idea can be naturally extended to corpus summarization by selecting a diverse set of documents that maximizes coverage. In this approach, every word in a document has a weight associated with it, indicating how important it is to cover this word in the summary. These document weights are either determined through a heuristic [10] or learned [19]. Documents are selected so that the total weight of the covered words is maximized; this is illustrated in Figure 1. In this example we want to select at most two documents out of three and each covered word has unit weight. We see that by selecting $d_1$ and $d_2$ we achieve the best score of the summary, since we cover the maximum number of words.

Formally, let $U = \{d_1, d_2, \ldots\}$ be the set of all documents in the corpus, where each document is represented as a bag of words.

**Algorithm 1** for greedy submodular function maximization.

$S^* \leftarrow \emptyset$
$A \leftarrow U = \{d_1, ...\}$
**while** $A \neq \emptyset$ **and** $|S^*| < k$ **do**
  $z \leftarrow \arg\max_{d \in A} F(S^* \cup \{d\}) - F(S^*)$
  $S^* \leftarrow S^* \cup \{z\}$
  $A \leftarrow A \backslash \{z\}$
**end while**

The word coverage objective function, for any $S \subseteq U$, is defined as follows:

$$F(S) = \sum_w \theta(w) \max_{d \in S} \phi(d, w), \quad (1)$$

where, $\phi(d, w)$ represents the weight of a word $w$ in the document $d$. One common choice of $\phi$ is the TFIDF score [17]. Moreover $\theta(w)$ is the weight for the word $w$ depending on our belief of the word's importance.

## 3.2 Optimization via Greedy Algorithm

With the objective function defining the score of a summarization $S$, corpus summaries are constructed by finding the set $S$ with the highest score $F(S)$. To obtain a summary we have to solve the following optimization problem:

$$S^* = \arg\max_{S \subseteq U} F(S). \quad (2)$$

An important property that enables the fast and accurate solution of this optimization problem lies in the structure of $F(S)$. It is well known that the coverage objective $F(S)$ is monotone (i.e. $|S'| \geq |S| \implies F(S') \geq F(S)$) and submodular [8].

DEFINITION 1. *Given a set $U$, a function $f : 2^U \rightarrow \mathbb{R}$ is submodular iff for all $u \in U$ and all sets $S$ and $T$ such that $S \subseteq T \subseteq U$,we have,*

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T).$$

Submodular functions have a property that says that adding $u$ to a subset $s$ of $t$ increases $f$ at least as much as adding it to $t$. While maximizing monotone sub-modular functions is NP-hard [11], it is known that the greedy Algorithm 1 achieves a $1 - 1/e$ approximation to the optimum solution for any linear budget constraint [10, 8]. Further, this algorithm provides a $1-1/e$ approximation for any monotone submodular scoring function. The algorithm starts with an empty summary. In each step, a document is added to the summary that results in the maximum relative increase of the objective. The algorithm terminates when the budget $k$ is reached.

# 4. CORPUS SUMMARIZATION

While submodular summarization approaches have been very successful for conventional summarization problems, we argue that corpus summarization should not only optimize coverage of information content, but also reflect which documents and authors were important in the development of the corpus. In particular, we aim to include influence between documents into the summarization objective.

To illustrate the difference between a conventional summarization problem and the type of corpus summarization we envision, consider a corpus consisting of research papers covering two decades of a particular field. In such a scenario, conventional word-based coverage approaches would pick several (non-redundant) survey papers, or papers that otherwise touch on a lot of different areas, since their union will tend to cover the largest subset of words in the corpus. However, while these survey papers are indeed a good summary of the content, this selection will not provide any information about how the corpus developed over time, what papers opened new areas of activity, and which authors influenced the direction of the field. In the subsequent sub-sections, we show how the conventional coverage-based approach can be extended to provide summaries that not only optimize information content, but also reflect influence and importance of individual documents and authors.

To achieve this goal, the remainder of this section shows how to (a) incorporate time into the summarization problems, (b) formulate the summarization objective in terms of influence, and (c) show how corpora can be summarized not only through landmark documents, but also through influential authors and key-phrases.

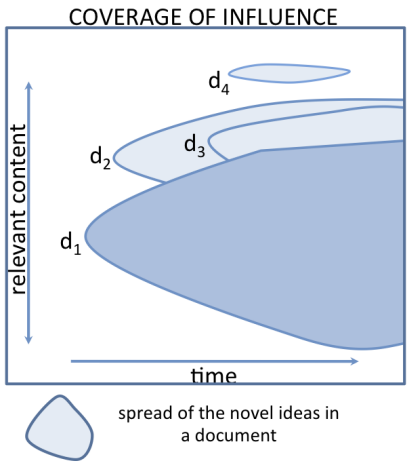


**Figure 2: Illustrating the coverage function for revealing influential documents.**

## 4.1 Summarization through Influential Documents

We now explore how the word-based coverage objective can be extended to summarize a corpus through a non-redundant set of influential documents. A pictorial illustration is shown in Figure 2, indicating how influential documents introduce ideas that increasingly cover the content of documents observed in later years. To get to an operational formalization of influence in the coverage model, we start with the following properties:

**Spread:** An influential document contains ideas that spread to other documents. The more an idea spreads, the greater is its influence. Note that this aspect of influence requires us to include a notion of time into the coverage objective, since ideas can only spread forward in time.

**Novelty:** A document should only be credited for generating influence with respect to some idea, if this idea was first proposed in that document. If an earlier document already contained that idea, influence should be credited to that earlier document. Note that this can be quite different from citation-based impact measures, which may credit influence to review papers or other papers that popularize an existing idea.

To capture *novelty* in the word-based coverage approach, we redefine $\phi(d, w)$ in the coverage objective (1). Intuitively, we do not

want to give document $d$ credit for an idea – as represented by its word distribution – if there already exist older documents $d'$, $t(d') < t(d)$ that already cover this idea. $t(d)$ denotes the year of publication of the document $d$. More formally, let $\mathcal{N}(d)$ denote the $k$-nearest neighbors (for example, based on cosine similarity) of the document $d$ among all the documents published before it, then we capture the novel contribution of a document as

$$\nu(d,w) = \max\left\{0, \min_{d' \in \mathcal{N}(d)} \{\phi(d,w) - \phi(d',w)\}\right\}. \quad (3)$$

In order to capture *spread* in the coverage objective, we enlarge the set of objects that need to be covered to word-time pairs $w_y$ for all words $w$ and years $y$. This allows the coverage objective to make a distinction between covering the word $w$ in a year $y$ and covering the same word in year $y'$. Formally, we generalize $\theta(w)$ in (1) and make it dependent on time, where each $\theta(w,y)$ separately defines how important word $w$ is for the given time $y$. We say that the importance of a word $w$ in year $y$ is determined by the sum of the TFIDF scores of the documents in year $y$:

$$\theta(w,y) = \sum_{d:t(d)=y} TFIDF(d,w). \quad (4)$$

Other weighting schemes can work here too (e.g. per-year inverse document frequencies) because our model is oblivious to the precise choice of $\theta$ (as long as it well represents the spread). Note that this allows us to model that some documents cover a word in certain years, but not in others. In particular, we say that a document $d$ only covers a word in those years that are later than its publication date $t(d)$. This allows us to formulate the objective for finding influential papers as follows:

$$F(S) = \sum_w \sum_y \theta(w,y) \max_{d \in S, y > t(d)} \nu(d,w). \quad (5)$$

The above objective multiplies the novel aspect of a word in a paper with how important a word is in the future years. Intuitively, the score is large when the set of selected documents $S$ contains documents with high novelty scores as well as a high influence in the future. We therefore maximize the above objective using the greedy Algorithm 1.

Note that this approach will not tend to select survey papers unlike the word coverage approach for two reasons. First (and most importantly), a survey paper will have a low novelty score since it is mostly based on previous work. Second, usually survey papers are written after a field is well developed, hence it does not cover all those documents that appeared before it in (5).

The key feature that differentiates our model from related coverage-based approaches is the insight that modelling the temporal context in which words appear (and not just the within-document context) can provide a strong signal for summarization tasks. We achieve this by not just using words as proxy for ideas in a coverage objective, but anchor words at specific time-points to gauge the influence. Additionally, modelling novelty helps us correctly attribute impact and avoids the pitfalls of citation-based impact measures.

## 4.2 Timelines of Document Influence

The previous section showed how the coverage objective can be extended to focus on influential papers, producing summaries that are organized by the publication date of the influential documents. However, dual to such summaries, we may also ask the following question: for each year $y$, what are the documents that most influenced the content of this year? This is illustrated in Figure 3.

In the following subsection, we formulate a coverage objective that identifies the $k$ documents that had the most influence in a year

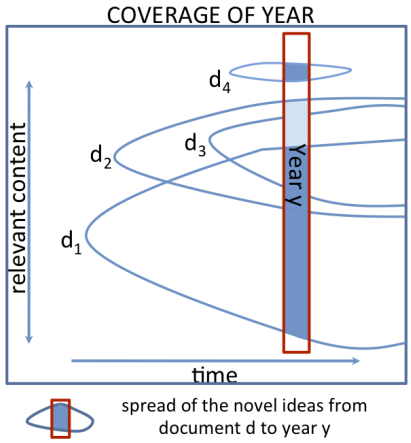


**Figure 3: Illustrating the influence of documents in a particular year.**

(for every year). Intuitively, a document $d$ influences a year $y$, if it was published before year $y$, i.e. $t(d) < y$, and the novel ideas from $d$ have substantial coverage in year $y$. Instead of selecting documents independent of year as in the previous section, we now allow our method to select an influential document to cover a particular year $y$. This means that our optimization problem now selects from a universe of document-year pairs $U_y = \{(d_i, y_i), ...\}$. Here $d_i$ is any document from the corpus, and $y_i$ is any year such that $y_i > t(d_i)$. This leads to the following objective which we seek to maximize.

$$F(S) = \sum_w \sum_y \theta(w,y) \max_{\substack{(d,\, y_d) \in S \\ y = y_d > t(d)}} \nu(d,w). \quad (6)$$

Similar to (5), the above objective multiplies the novelty score of word $w$ in a document $d$ with the importance of the word for a year $y$. However, (6) allows picking a different set of documents for each year $y_i$.

It is easy to see that $F(S)$ in (6) decomposes into a set of independent optimization problems – one for each year. We may therefore solve the following subproblem separately for each year and concatenate the solution for each year to obtain the solution of the original problem. Formally,

$$F_y(S_y) = \sum_w \theta(w,y) \max_{d \in S_y, y > t(d)} \nu(d,w). \quad (7)$$

Each of the above objectives is monotone submodular and can be solved using the greedy Algorithm 1.

## 4.3 Timelines of Author Influence

Analogous to selecting documents that had a large influence on a given year, we can also ask which authors were most influential. It is easy to extend the optimization problem from the previous section so that it selects influential authors. Denote with $d(a)$ the documents in the corpus that were authored by author $a$. The universe of items to select from now consists of author-year pairs $U_y = \{(a_i, y_i), ...\}$. Selecting an author $a_i$ for year $y_i$ implies that all documents the author wrote before year $y_i$ get selected. This leads to the following objective,

$$F(S) = \sum_w \sum_y \theta(w,y) \max_{\substack{(a,\, y_a) \in S \\ d \in d(a) \\ y = y_a > t(d)}} \nu(d,w). \quad (8)$$

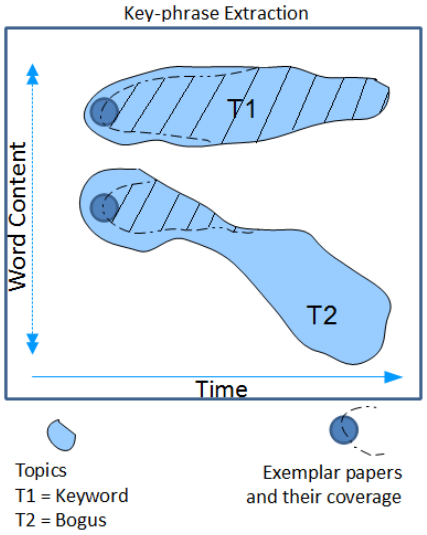


**Figure 4: Illustrating the difference in word distributions over time between a bogus term and a genuine keyword.**

which again can be broken into independent optimization problems for each year.

## 4.4 Summarizing Timelines with Key-Phrases

Summaries in terms of documents and authors still require the user to read through some documents from the collection. We now explore whether timelines of influence can be summarized through key-phrases. In particular, we aim to identify the points in time when new and influential ideas – as represented by a key-phrase – entered the collection.

While we already have operational definitions of novelty and influence, we still need to define what makes a key-phrase a good representative of an idea. We conjecture that a key-phrase that represents an idea well will be accompanied by stable word distribution over the years. For instance, documents that mention the phrase "HITS algorithm" will probably also mention several words related to that idea, whereas documents mentioning "Related work" need not have such a coherent set of overlapping words. The keyphrase T1 in Figure 4 is an example of a good key-phrase, since documents that contain T1 also share many other words. On the other hand, a key-phrase that is not a good representative of an idea may occur in documents talking about a variety of different ideas. T2 in Figure 4 is an example of a bad key-phrase.

We formalize this definition of key-phrases as follows. Define the universe of elements to choose from, $U = \{(p, y), \dots\}$, where $p$ is a candidate key-phrase and $y$ denotes the year when the key-phrase became influential. Let the subset of the corpus that mentions a candidate key-phrase $p$ be $D_p$. Intuitively, we wish to associate with $(p, y)$ a representative document $d^* \in D_p$ which was published in year $y$ and which was the most influential document in the subsequent development of $D_p$. According to our conjecture, for a bogus keyphrase, the associated $d^*$ will achieve very poor coverage of the word content observed in documents of $D_p$ that were published after $y$, while influential keyphrases will have a document that covers the associated stable word distribution very well. Following (4), we model the importance of covering a word

**Algorithm 2** for greedy submodular function maximization with budget constraint.

$S^* \leftarrow \emptyset$
$A \leftarrow U = \{p_1, ...\}$
$z^* \leftarrow \arg\max_{p \in A | C(p) < k} F(\{p\})$
**while** $A \neq \emptyset$ **and** $C(S^*) < k$ **do**
  $z \leftarrow \arg\max_{p \in A} \frac{F(S^* \cup \{p\}) - F(S^*)}{C(p)}$
  $S^* \leftarrow S^* \cup \{z\}$
  $A \leftarrow A \backslash \{z\}$
**end while**
**if** $F(S^*) < F(z^*)$ **then**
  **return** $z^*$
**else**
  **return** $S^*$
**end if**

in $D_p$ as $\theta(w, y)^p$. More precisely,

$$\theta(w, y)^p = \sum_{d \in D_p : t(d) = y} TFIDF(d, w).$$

$$d^*(p, y) = \arg\max_{d \in D_p : t(d) = y} \sum_w \sum_{y' > y} \theta(w, y')^p \nu(d, w).$$

With this $d^*$ for each element in $U$, we can formulate the objective

$$F(S) = \sum_w \sum_y \max_{\substack{(p_i, y_i) \in S \\ y = y_i}} \nu(d^*(p_i, y_i), w) \theta(w, y)^{p_i}. \quad (9)$$

Again, the objective decomposes into independent sub-problems for each year, and we can rewrite it for each year $y$ as,

$$F_y(S) = \sum_w \max_{(p_i, y_i) \in S, y_i = y} \nu(d^*(p_i, y_i), w) \theta(w, y)^{p_i}. \quad (10)$$

Unlike in the previous optimization problems, we now associate a cost $C(p, y) = |\{d \in D_p : t(d) = y\}|$ with each element of $S$. This is done to encourage associating a key-phrase with the point in time when it begins to gain popularity. The number of documents published in a year mentioning a key-phrase is used as a proxy for the maturity of an idea. The optimization problem is,

$$S_y{}^* = \arg\max_{S \subset U} F_y(S)$$
$$\text{subject to} \sum_{(p,y) \in S} C(p, y) \leq K.$$

This formulation is an instance of the budgeted coverage problem with a linear cost constraint, and the greedy Algorithm 2 is $(1 - 1/\sqrt{e})$ optimal [10, 8].

## 4.5 Alternate Formulation using Global Optimization and Adaptive Budget

For simplicity, so far we have decomposed the timeline-generating optimization problems into independent sub-problems, one for each year. This was possible since we imposed a cardinality constraint for each year. However, we can also define a global optimization problem across all years that constrains the maximum amount of content covered in each year. This results in a summary that will choose more documents from years that actually contain more interesting information. Formally, We change the global objective

from (6) into

$$F(S) = \sum_w \sum_y \min\{F_y(S_y), \tau F_y(U_y)\}, \qquad (11)$$

where the parameter $\tau$ determines how much relative word content per year we want to cover, $F_y(S_y)$ is as defined in (7) and $U_y$ is the whole universe for year $y$. This global $F(S)$ is also monotone submodular and can be solved using the greedy Algorithm 1. Whereas earlier we had a cardinality constraint parameter $k$ to set for each year, we have to set one global parameter $\tau$ (setting it higher results in more detailed summary) and one global $k$ (higher values result in longer summary) now.

# 5. EXPERIMENTS

In this section, we empirically evaluate our proposed models on publicly available datasets. We first describe the datasets and then present the results of our experiments along with evaluation metrics. The experimental results show the advantage of our approaches compared to other baselines in addition to good qualitative results.

## 5.1 Datasets

We used three corpora containing research publications for evaluating our proposed approaches. The Neural Information Processing Systems (NIPS) corpus contains 1955 published papers over a span of 14 years. Similarly, the Association for Computational Linguistics (ACL) corpus [15] contains 18041 papers published in a number of conferences over a span of 39 years. We also collected the set of papers published in the proceedings of SIGIR and CIKM conferences over the years available from CiteSeer. This corpus contains 2097 papers published over a span of 18 years (the last year being 2007). In all cases, we associate each document (paper) with its publication year and limit ourselves to 12 consecutive years ending at the year before last (the citation graph is also constrained only to those years). Since we compute novelty of a document based on the nearest neighbors from the past, we also used the year immediately before our subset for this purpose in the subsequent experiments. The last year is skipped because it does not have any citations from the future. We did not use the early years in ACL and SIGIR-CIKM corpus because they contain significantly less papers and citations compared to other years.

All datasets include citation graphs which we use for evaluation purposes, however our method does not require citation information and could thus be easily applied to other document collections. Note that the citation graphs are sparse as they do not include references to and from the papers outside the corpus. The NIPS collection has 1512, the ACL collection has 82892 and the SIGIR-CIKM collection has 1750 citations between papers inside the corpus. In addition to regular research papers, NIPS corpus also contains meta documents representing volume indices. We removed such documents manually since they are very easy to spot. We also pruned the words and retained only those words which occurred at least twice in a document and in at least three documents in the corpus. This simple heuristic removed a lot of noise introduced by the OCR system, and allowed us to meaningfully interpret influence. We represent every document by the TFIDF score (computed on the whole corpus) of the words contained in it after pruning, and normalize the resulting document vector to unit length. To compute the nearest neighbor in the past (to determine novelty), we use the cosine similarity between the document vectors. We do not require the exact nearest neighbors, and in case of a very large corpus, approximate methods to find similar documents can be employed to sidestep the quadratic time complexity of this step.

## 5.2 Influential Documents

The word coverage approach (from Section 3) obtains a summary by maximizing the word coverage. In Section 4.1 we argued that influential documents have novel ideas which subsequently spread through the corpus. In this experiment, we select the most influential papers based on our objective (5) which captures novelty of a document and its sphere of influence, and compare it with those selected by the simple word coverage objective (1). Since there is no standard way of measuring the influence of a paper, we resort to the citation structure available in the corpora. To quantitatively evaluate whether the selected papers were indeed influential, we compute the total citation count for the set of papers (i.e., the number of times these papers were cited by documents in the corpus) selected by any algorithm. There have been several criticisms of citation-based impact measures and some effort [5, 26] addresses them. However, for a comparative study, we believe citation counts are the least biased choice in this setup.

To provide a point of comparison for the coverage based approaches, we considered several baselines. The simplest baseline was to randomly select documents until the budget was reached (*random*). Another baseline that we considered was to select the most prolific authors in the conference (in terms of number of accepted papers) and then select the required number of papers from the union of their papers (*authors*). More concretely, we first rank the authors according to the number of papers in the collection they authored. Next, we pick the 10 most prolific authors. Finally, we sample uniformly at random with replacement from the set of authored papers the same number of papers for each author. We also computed the upper bound on the total citation count possible for a selection by selecting papers with the highest observed citation inlinks in the corpus (*bound*).

We selected 100 documents from the NIPS, SIGIR-CIKM and ACL corpus that maximize the respective objective function. Our experiments presented here are using unigrams as the elements in the universe. However, our methods can use other types of elements (e.g. bigrams formed from consecutive words, for which we observed a similar trend in the results). The results of this experiments are provided in Table 1. We also provide standard error of these results; they were estimated from the citation count on 10 re-runs of 70% subsampling of the corpus.

From the table, it is clear that our approach (*infl. papers*) gets significantly higher citations compared to the word-coverage approach (*word cover*) in all corpora. Moreover, finding influential papers is computationally cheap (with running time linear in the size of the corpus multiplied by the number of selected papers if we do not count the preprocessing step of computing nearest neighbors for novelty score) and, for example, takes a few seconds for the NIPS corpus on a standard desktop computer.

| method | NIPS | ACL | SIGIR-CIKM |
|---|---|---|---|
| Random | 64 (4.6) | 422 (41) | 84 (7.9) |
| Authors | 115 (4.0) | 1097 (49) | 86 (4.8) |
| Word Cover | 92 (2.7) | 799 (170) | 96 (5.2) |
| Infl. Papers | 196 (12.8) | 1842 (111) | 217 (14.5) |
| Bound | 521 (11.0) | 9787 (143) | 815 (13.5) |

**Table 1: Total citations obtained by the papers selected for influential documents and baselines using unigrams. All results use 1-NN for novelty score. The values in parentheses indicate standard error.**

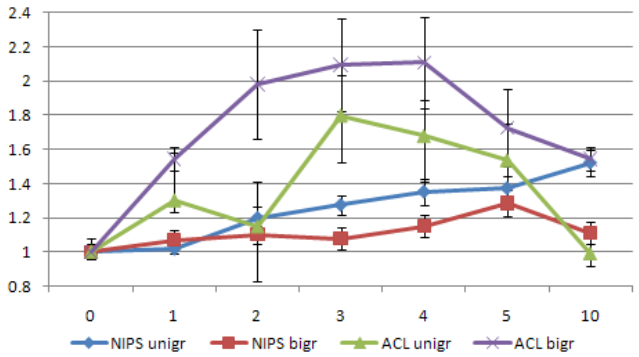


**Figure 5: Comparison of results of word coverage approach when using different values of $k$ (number of nearest neighbors for computing novelty score) on NIPS and ACL corpus for unigrams and bigrams. The horizontal axis represents the value of $k$ and the vertical axis relative performance (number of citations) when compared to not using the novelty score (i.e. $k = 0$).**

## 5.3 Impact of Using Novelty Score

Our approach uses the novelty score (introduced in Section 4.1, Eq. (3)) to credit a document for an idea only if it was the first one proposing it. Novelty is captured by considering $k$ nearest neighbors in the past and subtracting their word weights (clipping at 0 to prevent negative values) from the current document. In this subsection we explore the impact of choosing different values of parameter $k$.

Results for the word coverage approach on NIPS and ACL (as examples of two slightly different behaviors) using unigrams or bigrams as elements in the universe are presented in Figure 5 . We would expect word coverage to improve when using novelty scores because the coverage most likely does not choose the initial (highly cited) paper but some later one with better coverage (e.g. a derivative paper that also incorporates some other ideas). This intuition is confirmed by our results showing that using more neighbors improves the score as we incorporate more and more information about novelty. After a point we can see that performance starts dropping again because we are subtracting too much content.

Almost all coverage approaches benefit from using 1-NN, but increasing $k$ only improves performance for word coverage approach. We believe that using 1-NN helps because it mimics a language background model and penalizes frequent non-content words, while increasing $k$ above that does not bring significant benefits because we already model temporal behavior with the choice of our model.

## 5.4 Timelines of Document Influence

In this sub-section, we evaluate our approach to create timelines of document influence and compare it against several other baselines. For each NIPS, ACL and SIGIR-CIKM corpus, we select 10 documents per year.

Again, we considered the random baseline (*random*) and the 10 most prolific authors (*authors*). The authors baseline is constructed as follows: first create a union of all papers by 10 authors with the highest number of accepted papers, and for each year select 10 documents randomly from this union (with replacement) published on or before this year. We computed the upper bound on the citations (*bound*) by selecting papers with highest citation count in a given year (i.e. we count only citations occurring in that particular year) – we call this the *current citations*.

We evaluate the selections based on the citation network as before. In the previous section, our evaluation was based on the total number of citations a paper obtained. However, in this section, it is based on the current citations. To select a timeline of influential papers, we select papers that have maximum influence in a particular year (for each year). So, to quantitatively evaluate the selections, if a paper is selected as influential in the year $y$, we count the number of citations it gets in the year $y$ (i.e. only citations from papers citing it in this year count) and then sum them across all years.

Results for this experiment are summarized in Table 2. We can see that *random* baseline and *authors* have inferior performance compared to our approach (*timeline*). Note that the gap between our approach (*timeline*) and the *bound* is larger than in the influential papers experiment. We believe this is due to *timeline* being an inherently harder problem – not only do we have to find influential papers but we also have to specify exactly when were they influential (as the evaluation metric counts only citations from papers citing in that selected year). Our approach to constructing timelines is fast to compute (time complexity is linear in the number of years, papers selected and corpus size) and, e.g., takes less than 3 seconds on the NIPS corpus on a standard desktop computer.

| method | NIPS | ACL | SIGIR-CIKM |
|---|---|---|---|
| Random | 14 (1.4) | 85 (11) | 11 (1.2) |
| Authors | 14 (1.2) | 84 (14) | 7 (1.0) |
| Timeline | 60 (3.0) | 190 (14) | 30 (1.9) |
| Bound | 269 (3.0) | 3316 (11) | 367 (4.1) |

**Table 2: Current citations (i.e. number of citations from papers citing in that particular year) obtained by the papers selected for timeline and baselines using unigrams as elements of the universe. All results are for 10 re-runs of 70% subsampling and using 1-NN for novelty score. The values in parentheses indicate standard error.**

## 5.5 Timelines of Author Influence

In this section, we conducted experiments based on the objective proposed in Section 4.3 to select timelines of Author Influence. Instead of selecting papers we now consider meta-documents describing authors and construct a timeline showing which authors were important and when. In addition to this, for each selected author we constrain the corpus to that author's papers and find the most influential ones (including their timeframe of prominence).

We present a visualization of the results for NIPS corpus in Figure 6. By just looking at the plot it is easy to gain some insight into the development and content of the corpus. Features such as some authors having an influence throughout the whole corpus (e.g. Jordan, Sejnowski) are easy to spot. We can also see that some authors have had more influence only in specific timeframes (e.g. Hinton in the early years and Smola in the later years). In addition, looking at the selection of an author's most influential papers gives us insight into what topics they usually write about.

Although there are quantitative metrics which might be used to judge the output of the system to pick influential authors (e.g. *H-index*), note that we require these metrics to be computed for the timeframe of the collection only. Given the very sparse citation graph, using only observed intra-collection citations is expected to be a noisy signal. Qualitative results clearly indicate good performance of our approach and we feel that any simple adaptation of existing measures would not give a significantly better insight.

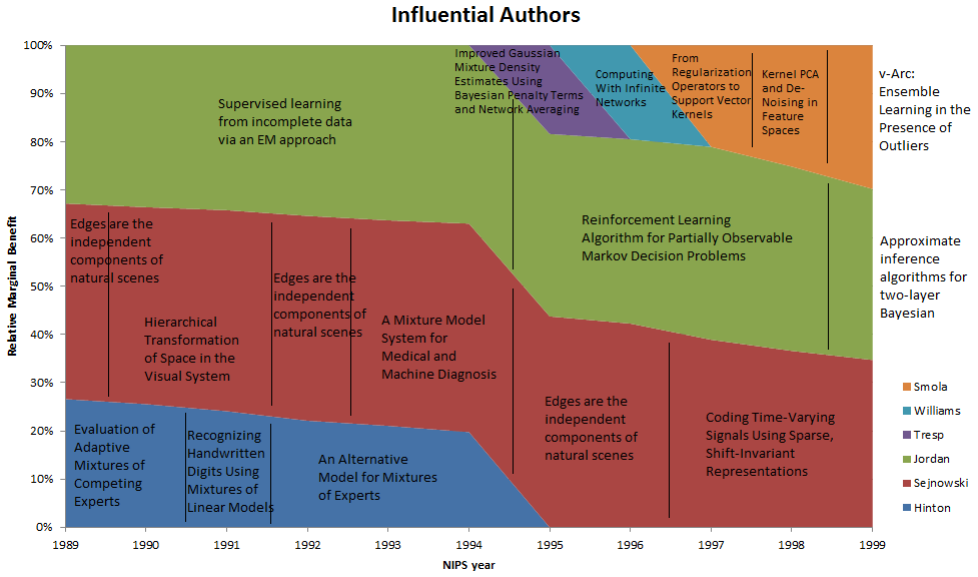


**Figure 6: An example of applying our framework to select the most important authors and the most important papers for the given authors using our framework. In each of 11 consecutive years of NIPS we selected three authors and then we selected that author's most influential paper in a particular year. The width of the author's slice (relative marginal benefit) represents the importance relative to other selected authors (which is computed as decrease in objective score if we remove that author from the collected set).**

## 5.6 Key-phrase Extraction

With the formulation described in 4.4, we ran experiments to find prominent key-phrases in each of the scientific corpora. For the set of candidate key-phrases, we considered trigrams and bigrams that occurred in at least $0.2\%$ of the documents in the respective corpus. Moreover, if a trigram is admitted to the set of candidate key-phrases, the constituent bigrams are not considered as candidates. This is a simple heuristic that recognizes that the lexical unit for phrases is usually a trigram or bigram and greedily prefers trigrams. More sophisticated ways to determine the set of candidates are possible, say independently running a Part-Of-Speech tagger and considering only noun phrases. The number of candidate key-phrases using this heuristic rule is $3035$ for the NIPS corpus, $8139$ for the SIGIR-CIKM corpus and $4687$ for the ACL corpus. The fewer number of candidates in the ACL corpus is explained by the fact that requiring the document frequency of bigrams or trigrams to be $0.2\%$ of a much larger corpus is a more restrictive filter.

We lack ground truth key-phrases to evaluate the output of our system; also, it is hard to quantitatively judge the quality of an influential key-phrase's associated timestamp. We therefore estimate the average citations in the collection of documents that mentions a key-phrase as a measure of its quality. Concretely, for a key-phrase $p$ and the subset of the corpus $D_p$ that mentions it,

$$Score(p) = \sum_{d \in D_p; d' \in D_p; d' \neq d} Cite(d \leftarrow d')/|D_p|$$

where $Cite(d \leftarrow d')$ indicates that document $d$ is cited by $d'$. We optimized the objective in Section 4.4 for each year with a budget of 3, and collected the set of all unique key-phrases. The reported scores for this approach (presented as *TimeCov* in Table 4) are the sum of $Score(p)$ for each unique collected key-phrase $p$. As a point of comparison, we also report the number of unique key-phrases collected as *Count*. A simple baseline for this experiment would be to pick the most frequent key-phrases occurring in the corpus in each year: this approach is hindered by the frequent occurrence of redundant phrases. For instance, "neural network" in the NIPS corpus,"natural language" in the ACL corpus and "information retrieval" in the SIGIR-CIKM corpus appear in such an overwhelming majority of documents over all the years as to drown out other informative candidates. This baseline is reported as *MostFreq* in our results. Another approach we compare with is to pick candidates that optimize the $Score$ directly in each year; this can be interpreted as an upper bound for this evaluation metric. We also provide the collected candidates from the coverage approach and one that optimizes the $Score(t)$ directly for the SIGIR-CIKM corpus in Table 3. Several informative phrases that come from diverse areas of research covered in SIGIR and CIKM get selected in the coverage approach. Furthermore, a visualization of the key-phrases over years for the NIPS corpus is shown in Fig. 7. Area of the shaded region corresponding to a term represents the fraction of documents observed in the corpus in that year that mention that particular term.

| | NIPS | | ACL | | SIGIR-CIKM | |
|---|---|---|---|---|---|---|
| Method | Count | Score | Count | Score | Count | Score |
| MostFreq | 13 | 2.13 | 20 | 34.33 | 20 | 7.41 |
| TimeCov | 17 | 4.68 | 77 | 116.92 | 29 | 12.28 |
| Bound | 13 | 6.09 | 96 | 124.10 | 29 | 18.82 |

**Table 4: Quantitative results of keyphrase extraction.**

# 6. CONCLUSIONS

This paper presented a submodular framework for temporal corpus summarization. We extended the notion of word coverage and asserted that summaries cover important concepts by covering associated words over a time interval. A timeline of influential documents, or authors, or coherent key phrases was constructed using

| Year | Coverage | | CiteScore | |
|---|---|---|---|---|
| | Keyword | Marginal Influence | Keyword | Marginal Influence |
| 1995 | relevant document | 17.240 | information retrieval | 20.0 |
| | query expansion | 5.464 | singular value | 5.0 |
| | | | training set | 5.0 |
| 1996 | search engine | 8.228 | speech recognition | 3.5 |
| | semantic indexing lsi | 6.034 | block size | 2.0 |
| | filtering system | 2.818 | | |
| 1997 | web search | 11.853 | retrieval system | 5.0 |
| | training data | 7.871 | test collection | 5.0 |
| | language model | 6.760 | | |
| 1998 | language model | 5.411 | summarization system | 8.0 |
| | retrieval model | 4.799 | naive bayes | 7.0 |
| | learning algorithm | 4.024 | unjudged document | 6.0 |
| 1999 | language model | 12.358 | general english | 10.0 |
| | | | pearson correlation | 5.0 |
| 2000 | clustering result | 4.354 | cumulative gain | 8.0 |
| | document model | 3.130 | expansion term | 4.0 |
| | cori algorithm | 2.645 | event detection | 4.0 |
| 2001 | cross-language information | 4.897 | smoothing method | 14.5 |
| | user information | 3.450 | topic distillation | 7.0 |
| | hits algorithm | 3.044 | | |
| 2002 | training example | 2.099 | translation disambiguation | 3.0 |
| | term dependency | 1.868 | hoc retrieval | 3.0 |
| | input stream | 1.698 | | |
| 2003 | query language | 2.509 | image feature | 5.0 |
| | feature selection | 1.779 | finding expert | 5.0 |
| | document clustering | 1.751 | novelty detection | 4.0 |
| 2004 | training image | 1.887 | regularized logistic regression | 3.0 |
| | inverted index | 1.413 | label information | 3.0 |
| | element retrieval | 1.100 | web browser | 3.0 |
| 2005 | xml retrieval | 3.926 | existing retrieval function | 3.0 |
| | | | new system | 2.0 |
| | | | index construction | 2.0 |

**Table 3: The list of key-phrases for SIGIR-CIKM selected by the greedy algorithm solving the budgeted coverage problem with budget of 3 and by optimizing the citation score.**

our approach, providing concrete suggestions for further and more detailed exploration of the corpus contents. Our approach leveraged both the novelty of a document as well as its influence in the development of the corpus and relied only on word features; in particular, it does not require a citation structure to infer influence across time. Therefore it is applicable to any textual collection which provides timestamped documents. Our optimization objectives used monotone submodular functions to trade-off relevance and redundancy elegantly, and were solved using an efficient greedy algorithm with a constant factor approximation guarantee. We empirically demonstrated that our approach performs better than several baselines using citation based performance measures and provided qualitative timelines for a few scientific corpora.

## 7. ACKNOWLEDGEMENTS

This research was funded in part through NSF Awards IIS-0812091, IIS-0905467, and IIS-1217686.

## 8. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR*, pages 10–18, New York, NY, USA, 2001. ACM.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, New York, NY, USA, 1998. ACM.

[4] C. C. Chen and M. C. Chen. Tscan: a novel method for topic summarization and content anatomy. In *SIGIR*, pages 579–586, New York, NY, USA, 2008. ACM.

[5] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8 – 15, 2007.

[6] K. El-Arini and C. Guestrin. Beyond keyword search: discovering relevant scientific literature. In *KDD*, pages 439–447, New York, NY, USA, 2011. ACM.

[7] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD*, pages 289–298, New York, NY, USA, 2009. ACM.

[8] S. Khuller, A. Moss, and J. S. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1), 1999.

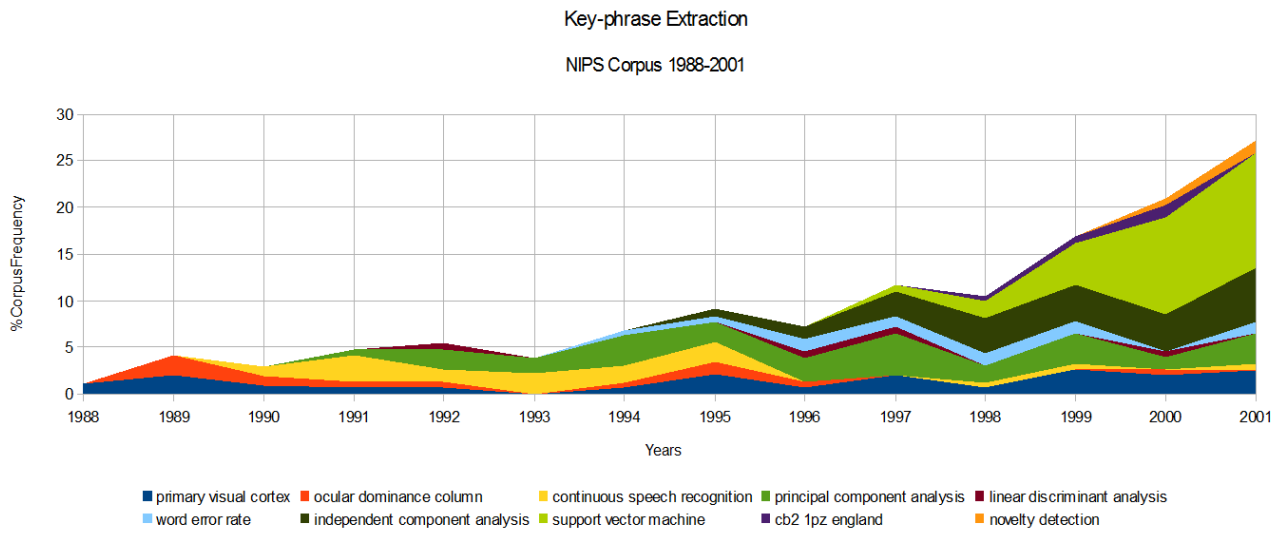


**Figure 7: A timeline showing the evolution of the key-phrases selected by the coverage approach in the NIPS corpus.**


[9] J.-M. Lim, I.-S. Kang, J.-H. Bae, and J.-H. Lee. Sentence extraction using time features in multi-document summarization. In *Information Retrieval Technology*, volume 3411 of *Lecture Notes in Computer Science*, pages 82–93. Springer Berlin / Heidelberg, 2005.

[10] H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *HLT*, pages 912–920, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] R. McDonald. A study of global inference algorithms. In *Lecture Notes in Computer Science*, 2007.

[12] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM*, pages 446–453, New York, NY, USA, 2004. ACM.

[13] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[14] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.

[15] D. R. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL anthology network corpus. In *Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore, 2009.

[16] K. Raman, T. Joachims, and P. Shivaswamy. Structured learning of two-level dynamic rankings. In *CIKM*, 2011.

[17] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA, 1987.

[18] B. Shaparenko and T. Joachims. Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases. In *KDD*, pages 619–628, 2007.

[19] R. Sipos, P. Shivaswamy, and T. Joachims. Large-margin learning of submodular summarization methods. In *EACL*, 2012.

[20] I. Subašić and B. Berendt. From bursty patterns to bursty facts: The effectiveness of temporal text mining for news. In *ECAI*, pages 517–522, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.

[21] A. Swaminthan, C. Metthew, and D. Kirovski. Essential pages. In *Technical Report, MSR-TR-2008-15*, Microsoft Research, 2008.

[22] R. Swan and J. Allan. Automatic generation of overview timelines. In *SIGIR*, pages 49–56, New York, NY, USA, 2000. ACM.

[23] H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *EACL*, pages 781–789, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[24] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl. Enhancing digital libraries with techlens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '04, pages 228–236, New York, NY, USA, 2004. ACM.

[25] M. Wu, W. Li, Q. Lu, and K.-F. Wong. Event-based summarization using time features. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '07, pages 563–574, Berlin, Heidelberg, 2007. Springer-Verlag.

[26] E. Yan and Y. Ding. Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8):1635–1643, 2010.

[27] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *SIGIR*, pages 745–754, New York, NY, USA, 2011. ACM.

[28] Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *ICML*, pages 271–278, 2008.