

# Discovering Saliency in Textual Elements using Graph Mutual Reinforcement SI508 Project

Ahmed Hassan  
*hassanam@umich.edu*

## Abstract

The problem of identifying the most salient terms and/or sentences from a set of documents has gained great interest in recent years. Identifying the set of the most salient terms in a set of documents is usually called automatic keyword extraction or terminology extraction. Extracting the most salient set of sentences from a document or a set of documents is used for extractive summarization where a summary is generated by selecting a smaller set of sentences that best represent the information content in the documents. In this work, we utilize the duality of the important term extraction and important sentence extraction problems. This duality allows us to present a single graph based method for the automatic extraction of both salient terms and salient sentences from a collection of documents. The approach employs graph mutual reinforcement to rank terms and sentences according to their importance with respect to the main topic of the documents collection.

## 1 Introduction

Recent years have witnessed an increasing interest in graph based methods for natural language processing. In this work we present and investigate a new graph based approach for identifying the most salient words and sentences in a set of documents. The approach depends on constructing a bipartite graph of words and sentences, and deploys graph based mutual reinforcement to weight the importance of these words and sentences. The mutual reinforcement is used to automatically identify the most informative sentences/words, where words that belong to several important sentences tend to be more important. Also, sentences that contain several important words tend to be important. The intuition is that such document collections are redundant, i.e. different information could be found many times in different documents and by different representation. The problem can therefore be seen as hubs (words) and authorities

(sentences) problem which can be solved using the Hypertext Induced Topic Selection (HITS) algorithm [7]. HITS is an algorithmic formulation of the notion of authority in web pages link analysis, based on a relationship between a set of relevant “authoritative pages” and a set of “hub pages”. The HITS algorithm benefits from the following observation: when a page (hub) links to another page (authority), the former confers authority over the latter. By analogy to the authoritative web pages problem, we could represent the sentences as authorities and words as hubs, and use mutual reinforcement between sentences and words to weight the most authoritative sentences and words. The rest of this report proceeds as follows: in Section 2 we discuss previous work followed by a brief definition of our general notation in Section 3. A detailed description of the proposed approach then follows in Section 4. Section 5 discusses experimental results while the conclusion is presented in Section 6.

## 2 Related Work

The work presented in this report can fall under two different research problems: Terminology extraction, and Extractive summarization.

Terminology extraction, or term extraction, is a sub-task of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given set of documents.

Several approaches have been presented for extracting important keywords or terminology from text. Some of them use standard supervised classification techniques [3]. Others try to utilize linguistic resources [6].

Extractive summarization is the task of summarizing a document or a set of documents by identifying and extracting the most important or representative sentences in them.

Previous work on extractive summarization used techniques such as graph matching [1], maximal marginal relevance[9], or language generation[11].

Several graph based techniques have been proposed for extractive summarization. One of the first attempts is using degree centrality in single document text summarization [14]. In this approach, degree scores are used to extract the important paragraphs of a text. Another way of identifying centrality of a sentence is often defined in terms of the centrality of the words that it contains. A common way of assessing word centrality is to look at the centroid of the document cluster in a vector space[12, 13]. [2] uses random walks on sentence-based graphs to identify most important sentences in a set of documents.

Graph based mutual reinforcement using HITS has been used in different areas other than web page ranking. [8] present an approach to improving the precision of an initial document ranking by performing re-ranking based on centrality within bipartite graphs of documents (on one side) and clusters (on the

other side). Similar idea was used for unsupervised [4] and semi supervised [5] extraction of relation between entities in text.

### 3 Background

In graph theory, a graph is a set of objects called vertices joined by links called edges. A bipartite graph, also called a bi-graph, is a special graph where the set of vertices can be divided into two disjoint sets with no two vertices of the same set sharing an edge.

Hypertext Induced Topic Selection (HITS) is a link analysis algorithm for rating, and therefore ranking, web pages using their authority and hub values. Authority value estimates the value of the content of the page; hub value estimates the value of its links to other pages. These values can be used to rank Web search results. The HITS algorithm makes use of the following observation: when a page (hub) links to another page (authority), the former confers authority over the latter. HITS uses two values for each page, the "authority value" and the "hub value". "Authority values" and "hub values" are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that authority. A hub value is the sum of the scaled authority values of the authorities it points to.

HITS was originally proposed as a mechanism of ranking web search results. It starts with a query represented by a query string. It uses that query to construct a subgraph of the web graph that is relevant to the search query. Standard text retrieval techniques are used to find web pages that contain the query string. This set of web pages is called the root set. The subgraph, also called a focused graph, is obtained by extending the root set by adding any web pages along the edges entering or leaving the root set. HITS link analysis is then employed to assign two values to each page, authority and hub values.

HITS is different from other web page ranking algorithms, like PageRank [10], in several aspects. First It is processed on a small subset of relevant documents (i.e the focused graph), not all documents. Second, It is executed at query time, and not at indexing time. Finally, It computes two scores per document (hub and authority) as opposed to a single score.

The last difference of those mentioned above is of major importance to our usage of HITS for solving the important term/sentence extraction problem. When used with we web page search, HITS is employed in a symmetric settings. By symmetric we mean that the set of graph nodes acting as hubs is the same as the set of graph nodes acting as authorities. Actually, each web page is represented by a two nodes in the bipartite graph, one representing the web page when acting as a hub and the other representing the web page when acting as an authority. This different roles in the graph suggest that we can employ HITS in an asymmetric setting, where the set of hub nodes is different from the set

of authority nodes. More details about how we used that idea to tackle our problem will be outlined in the following sections.

## 4 Approach

The approach we propose depends on the construction of a bipartite graph connecting terms and sentences from a set of documents. Terms and sentences are then assigned ranks based on graph based mutual reinforcement techniques. We try to model the duality in terms and sentences relation which could be stated that important sentences usually contain important words and important words tend to occur in important sentences. The proposed approach is composed of two main steps namely, term/sentence graph construction and terms/sentences ranking. Both steps are detailed in the next subsections.

### 4.1 Graph Construction

The process of constructing the term/sentence graph starts with a set of documents and builds a bipartite graph relating terms and sentences in this set of documents. To accomplish this we have to select the terms and sentences to be included as graph node and propose a criteria for connecting any two nodes in the graph.

Let us start with sentences as it is much simpler. All sentences in the set of documents are mapped to nodes in the graph. Documents are decomposed into a set of sentences based on a simple procedure that considers each sequence of tokens ending with a fullstop or a new line as a sentence.

Choosing terms to include in the graph is more difficult. The most basic way is to tokenize the text in all documents into words and include all words as nodes in the graph. This technique has a serious drawback because it allows frequent words that occur in several sentences and carry no information content to be included in the graph. For example words like “the”, “to”,...etc does not contribute to the meaning of any sentence and can occur in several sentences. Before considering words to include in the graph, the text in all documents go through the following list of preprocessing steps:

- Tokenization: the process of converting a sequence of characters into a sequence of tokens. tokens can be roughly mapped to words. This process simply separates words from punctuation and outputs a sequence of tokens.
- Stemming: the process for reducing inflected (or sometimes derived) words to their stem, base or root form. This process help compress the space of terms by joining terms that has the same root or stem. For example the stemmer should identify the string "cats" and “cat” as based on the root

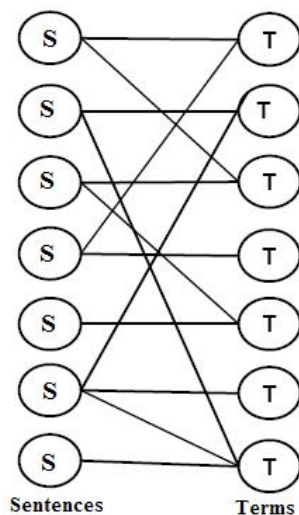


Figure 1: A bipartite graph representing sentences and terms

"cat", and "stemmer", "stemming", "stemmed" as based on "stem" and so on.

- Stopword removal: stopwords or non content words are words that do not contribute to the semantics of the sentence. The removal of this words usually help NLP algorithm to focus on content words only. Several lists of those such words exist. stopwords can be manually identified and can also be extracted from a large corpus by finding terms with low term frequency - inverse document frequency (TF-IDF) measure.

After applying the steps above, we add a node to the graph representing each term of the remaining terms.

The second step in graph construction is adding edges between terms and sentences. We scan all terms and sentences adding an edge between a word  $w$  and a sentence  $s$  if  $w$  occurs in  $s$ .

## 4.2 Ranking

The inherent duality in the sentences and terms relation suggests that the problem could be interpreted as a hub authority problem. This problem could be solved by applying the HITS algorithm to iteratively assign authority and hub scores to sentences and terms respectively.

Sentences and terms are represented by a bipartite graph as illustrated in figure 1. Each sentence or term is represented by a node in the graph. An edge connects a sentence and a term if the term occurs in the sentence. The sentence/term induction problem can be formulated as follows: Given a set of documents  $D$  containing a large set of sentences  $S$  and a large set of terms  $T$ , the problem is to identify  $S'$ , the most salient set of sentences and  $T'$  the most salient set of terms. The intuition is that the terms that occur in many different important sentences tend to be important and sentences containing many different important terms tend to be important. In other words; we want to choose, among the large space of sentences and terms in the data, the most informative, or “authoritative” sentences/terms. However, both  $S'$  and  $T'$  are unknown. The induction process proceeds as follows: each pattern  $s$  in  $S$  is associated with a numerical authority weight  $a_s$  which expresses how many terms occur in that sentence. Similarly, each term  $t$  in  $T$  has a numerical hub weight  $h_t$  which expresses how many sentences contained this term. The weights are calculated iteratively as follows:

$$a^{(i+1)}(s) = \sum_{u=1}^{T(s)} \frac{h^{(i)}(u)}{H^{(i)}}$$

$$h^{(i+1)}(t) = \sum_{u=1}^{S(t)} \frac{a^{(i)}(u)}{A^{(i)}}$$

where  $T(s)$  is the set of terms that occurred in sentence  $s$ ,  $S(t)$  is the set of sentences containing term  $t$ ,  $a^{(i+1)}(s)$  is the authoritative weight of sentence  $s$  at iteration  $(i+1)$ , and  $h^{(i+1)}(t)$  is the hub weight of term  $t$  at iteration  $(i+1)$ .  $H^{(i)}$  and  $A^{(i)}$  are normalization factors defined as:

$$H^{(i)} = \sum_{s=1}^{|S|} \sum_{u=1}^{T(s)} h^{(i)}(u)$$

$$A^{(i)} = \sum_{t=1}^{|T|} \sum_{u=1}^{S(t)} a^{(i)}(u)$$

### 4.3 Adding Weights to Arcs

Another refinement to the ranking process adds weights to arcs between sentences and terms. We used the TF-IDF for each term in the corresponding document to weight edges.

The tf-idf weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [14]. The term frequency in the given document is simply the number of times a given term appears in that document. The inverse document frequency is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient).

The weights are calculated iteratively as follows: $w_{us}$

$$a^{(i+1)}(s) = \sum_{u=1}^{T(s)} \frac{w_{us} * h^{(i)}(u)}{H^{(i)}}$$

$$h^{(i+1)}(t) = \sum_{u=1}^{S(t)} \frac{w_{us} * a^{(i)}(u)}{A^{(i)}}$$

where  $T(s)$  is the set of terms that occurred in sentence  $s$ ,  $S(t)$  is the set of sentences containing term  $t$ ,  $a^{(i+1)}(s)$  is the authoritative weight of sentence  $s$  at iteration  $(i + 1)$ , and  $h^{(i+1)}(t)$  is the hub weight of term  $t$  at iteration  $(i + 1)$ .  $H^{(i)}$  and  $A^{(i)}$  are normalization factors defined as:

$$H^{(i)} = \sum_{s=1}^{|S|} \sum_{u=1}^{T(s)} w_{us} * h^{(i)}(u)$$

$$A^{(i)} = \sum_{t=1}^{|T|} \sum_{u=1}^{S(t)} w_{us} * a^{(i)}(u)$$

## 5 Results

We implemented a Perl/C++ pipeline that takes any set of related documents and ranks terms and sentences according to their importance. The graph construction step is done using Perl, while the implementation of the HITS algorithm that assigns weights to different nodes is in C++.

We applied the approach to several sets of related documents. We report here a description of every set along with the highest/lowest ranked terms and sentences as extracted by our approach.

Topic	# Sents	# Terms	# Nodes	# Edges	Avg. Degree
Topic 1	80	126	206	1197	11.6
Topic 2	136	578	714	2703	7.6
Topic 3	56	302	358	811	4.5

Table 1: Some statistics about the three test topics.

The source of all documents was mainly different news streams from different sources. Documents of the first topic discuss the nomination and awarding of Nobel prize to former vice president Al Gore. The second set of documents is commenting on the decision of putting the mummy of the Pharaoh king Tutankhamun in public display for the first time since its discovery several decades ago. The last set of documents is talking about arresting several Europeans on Chad and the efforts of French president Sarkozy for releasing them. Some statistics about the three topics and their corresponding networks are stated in Table 1.

The difference between the weighted and unweighted cases was not significant with respect to our test data. However, we believe that further more objective testing would show an improvement in the case of using weighted edges. The highest/lowest ranked terms and sentences as extracted by our approach are reported below:

### Topic 1 (Al Gore Nobel Prize)

Top ranked sentences for are:

- OSLO , Norway - Former Vice President Al Gore was nominated for the 2007 Nobel Peace Prize for his wide-reaching efforts to draw the world’s attention to the dangers of global warming.
- Former U.S. vice-president Al Gore said he will use the recognition of winning the 2007 Nobel Peace Prize to change the world ’s consciousness about the challenges of global warming.
- The Nobel committee said that in awarding the peace prize to IPCC and Gore , it hoped to draw attention to the issue of climate change and the threat it poses to the future security of mankind.
- The Nobel committee said that in awarding the peace prize to IPCC and Gore , it hoped to draw attention to the issue of climate change and the threat it poses to the future security of mankind .
- The Norwegian Nobel committee announced the award Friday in Oslo, saying in a written statement that Gore and the UN Intergovernmental Panel on Climate Change have worked tirelessly to disseminate greater

knowledge about man-made climate change and lay the foundations for the measures that are needed to counteract such change.

Bottom ranked sentences for topic 1 are:

- The organization involves hundreds of scientists working to collate and evaluate the work of thousands more.
- It truly is a planetary emergency and we have to respond quickly.
- It points to the growing concern globally in what man is doing to the environment.
- I 'm going back to work right now . This is just the beginning .
- Story continues below.

Top ranked terms are:

- Nobel, Climate, prize, change, said

Bottom ranked terms are

- hundreds,news, going, back, right, individual

## **Topic 2 Tutankhamun's Mummy Public Display**

Top ranked sentences are:

- Egypt is set to put King Tutankhamun's mummy on public display for the first time.
- The mummy had to be reconstructed after Carter cut it into 18 pieces in order to gain access to amulets and other jewelery , said Mustafa Wazery , director of the Valley of the Kings .
- Tutankhamun has captured the world 's imagination in the decades since his 3,000-year-old mummy was found .
- The true face of ancient Egypt 's boy king Tutankhamun was revealed on Sunday to the public for the first time since he died in mysterious circumstances more than 3,000 years ago .
- The mummy risked being reduced to dust because of the rising levels of humidity due to the visitors , said Hawass , who heads the Supreme Council of Antiquities.

Bottom ranked sentences are:

- VOA 's Challiss McDonough is in Luxor and has this report.
- The host of gold and treasure found when his tomb was opened in 1922 has made him one of the most famous of ancient Egypt 's rulers .
- The pharaoh Akhenaton the Heretic was thought to have fathered Tutankhamun , but the identity of his mother is not known for sure .
- Every day hundreds of visitors file through his tomb in the Valley of the Kings on the west bank of the Nile in the southern Egyptian city of Luxor.
- Everyone is dreaming of what he looks like

Top ranked terms are:

- mummy, king, Tutankhamun, display, Egypt

Bottom ranked terms are:

- access, actual, addressed, agreed, amazing

### Topic 3

Top ranked sentences are:

- Three French journalists and four Spanish flight attendants detained in Chad over an alleged illegal attempt by a charity to fly African children to Europe were released yesterday after the French president , Nicolas Sarkozy , flew to Chad .
- Three French journalists and four Spanish flight attendants detained in Chad over an alleged illegal attempt by a charity to fly African children to Europe were released yesterday after the French president , Nicolas Sarkozy , flew to Chad .
- The Europeans , including nine French citizens , were arrested on 25 October when a charity called Zoe 's Ark was prevented from flying the children from eastern Chad to Europe , where the group said it intended to place them with host families .
- Six members of French group Zoe 's Ark are charged with fraud and abduction . Three members of a Spanish air crew are charged as accessories , as is a Belgian pilot who was arrested later

- Jean-Bernard Padare , a lawyer acting for the detainees , said those freed were three French journalists and four flight attendants from Spain

Bottom ranked sentences are:

- Deby said there was never a question about refusing the arrival of the European forces .
- France , the former colonial power , has troops stationed in Chad and will provide about half of up to 3,000 European Union forces
- We 're very relieved , journalist Marc Garmirian told reporters at the French airport .
- In remarks broadcast by France 's LCI television .
- This story has been viewed 114 times.

Top ranked terms are:

- Sarkozy, French, Chad, children, charity.

Bottom ranked terms are:

- acting, accessories, agency, arrival, air.

## 6 Conclusion

In this work, we tried to present a solution for the problem of identifying the most salient terms and/or sentences from a set of documents. We tried to exploit the duality inherent in the problems of extracting important terms and important sentences from text. This duality allows us to present a single graph based method for the automatic extraction of both salient terms and salient sentences from a collection of documents. The approach employs graph mutual reinforcement to rank terms and sentences according to their importance with respect to the main topic of the documents collection. The approach depends on constructing a bipartite graph of terms and sentences, and deploys graph based mutual reinforcement to determine the importance of these terms and sentences using the observation that important terms usually occur in important sentences, and important sentences usually contain important terms. The proposed approach achieved very promising results when tested on several documents sets from different news feeds.

## References

- [1] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 335–336, 1990.
- [2] G. Erkan and D. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.
- [3] J. Fernandez, A. Serrano, P. Martinez, and J. Villena. Automatic keyword extraction for news finder. *Adaptive Multimedia Retrieval, Lecture Notes in Computer Science*, 2004.
- [4] H. Hassan, A. Hassan, and O. Emam. Unsupervised information extraction approach using graph mutual reinforcement. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 2006.
- [5] H. Hassan, A. Hassan, and S. Noeman. A graph based semi-supervised approach for information extraction. In *Proceedings of the TexGraphs workshop, Human Language Technologies and North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, 2006.
- [6] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003.
- [7] J. Kleinberg. Authoritative sources in a hyperlinked environment. *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [8] O. Kurland and L. Lee. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of SIGIR*, pages 83–90, 2006.
- [9] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1, 2000.
- [10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [11] D. Radev, V. Hatzivassiloglou, and K. McKeown. description of the cidr system as used for tdt-2. In DARPA broadcast news workshop, 1999.
- [12] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 1, 2004.

- [13] D. Radev, A. Winkel, and M. Topper. Multi document centroid-based text summarization. In *ACL 2002 (Demo Session)*, 2002.
- [14] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing and Management*, 33, 1997.