**Connecting the Dots Between News Articles**

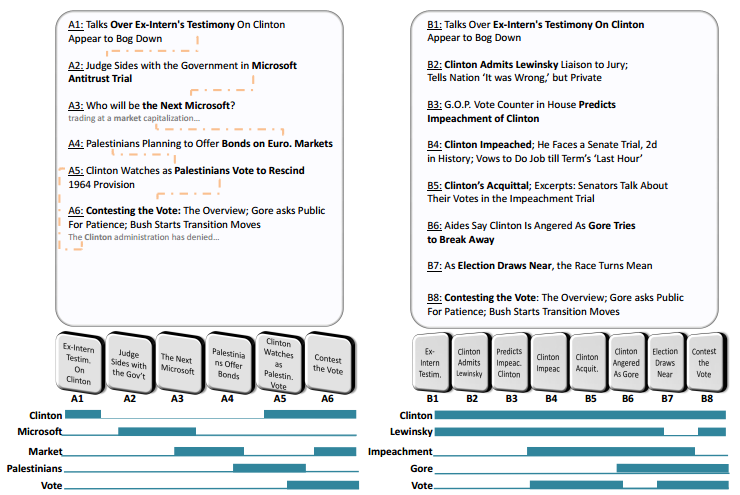
**DafnaShahaf, Carlos Guestrin**

The aim of the author is to build a chain of documents to connect two documents which were published at different times. The input to the algorithm is a set of documents with their timestamps, a start document(*s*) and an end document(*t*). The given set of documents can be chronologically ordered using their timestamps. Let there be *n* documents between *s* and *t.* There are 2n possible chains between s and t. There is need of scoring mechanism to choose the best possible chain.

Author has proposed two metrics for scoring a chain – *coherency* and *influence*. A coherent chain has a global coherent theme across the storyline i.e. all the documents in the chain belong to a single theme. Two consecutive documents in a chain has high influence(for a given word w) if the two documents are highly connected and w is important for connectivity.

The steps involved in the algorithm are as follows –

An example of coherent chain vs incoherent chain is given below.



In the example, the left chain is erratic and the documents are not logically connected. The right chain has highly connected documents and thus is more coherent.

Let D be a set of articles, and W a set of features (typicallywords or phrases). Each article is a subset of W. Given a chain (d1, ...,dn) of articles from D. An intuitive way to form a coherent chain is that every time a word appears in two consecutive documents we score a point.



Thus similar documents are placed next to each other. But it has 4 drawbacks – weak links, missing words, importance, jitteriness.

**Weak links:** A chain can have high coherence is most of the links are strong while few links are weak. Summing over the transitions can lead to ‘broken’ chains (having weak links). A more reasonable way is to consider the minimal transition score instead of the sum.



**Missing Links**: There are cases in which some words donot appear in an article, although they should have.For example, if a document contains ‘lawyer’ and ‘court’but not ‘prosecution’, chances are ‘prosecution’ is stilla highly-relevant word. Considering only words from the article can be misleading in such cases.

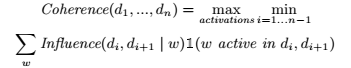
**Importance:**Some words are more important than others,both on a corpus level and on a document level. Two documents can have numerous common words but more important words must have influence on the transition between the two documents.

To address above two problems, the author suggests the concept of influence of di on di+1 through the word w. The calculation of Influence(di,di+1| w) has been discussed later. Intuitively, Influence(di , dj | w) is highif (1) the two documents are highly connected, and (2) w isimportant for the connectivity.



**Jitteriness:** Jitteriness is appearance and disappearance of patterns throughout the chain.

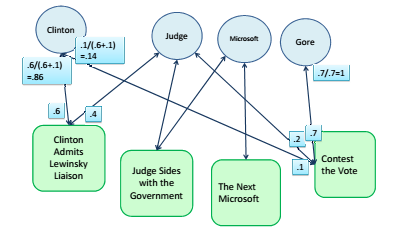
One way to avoid jitteriness is to consider the longest continuous stretch of each word. But words can have high influence on a transition even if they do not appear in the documents.The author defines an activation pattern arbitrarily for each word,and compute the objective based on it. The coherence isthen defined as the score under the best activation pattern.



Word acitvations can be binary or continuous. The author has considered word activation in the range [0,1]. The above mentioned equation is called ‘**Objective function**’.

**Calculating *Influence(di,dj | w)***

Construct a bipartite graph, G = (V,E), where V = VD U VW, VD is set of documents, VW is set of words. Edge weights represent the strength of the correlation between a document and a word. Author has used tool www.copernic.comto assign importance to each word and use these weights for document-to-word edges. Alternatively, TF-IDF weights can also be used. Since weights are interpreted as random walk probabilities, they are normalized over all words in thedocument.



Intuitively, if the two documents are connected, a short random walk starting from di should reach dj frequently. The stationary distribution is the fraction of the time the walker spends on each node:



Where Πi (v) is the stationary distribution of random walk starting from di. P (v | u) is the probability of reaching v from u and ϵ is random restart probability.

Let Πwi (v) be the stationary distribution for graph which has as a sink node. If w was influential, the stationary distribution of dj would decrease a lot The influence on dj w.r.t. w it defined as the difference between these two distributions, **Πi (dj ) – Πwi (dj ).**