Determining the Evolutionary Sequence of Versions of a Text with the Minimum Spanning Tree Algorithm

Katia P. Mayfield **Emanuel S. Grant** **Crystal Alberts**
Department of Computer Science **Department of English**
University of North Dakota
Grand Forks, ND 58201

Abstract

A new area in academics that directly influences the publishing industry is Digital Humanities. This area results of an increasing collaboration between traditional Humanities and computing technology. Two main topics in this field are digitization of data, whether that is printed literature or images of an artifact, and information that is considered to be "digital born". Literary analysis of digitized and "born digital" texts are becoming more common. In such analysis, one important factor is the progression of the text, which may not be well defined. This study proposes the use of a minimum spanning tree algorithm to predict an evolutionary path of the literary work, estimating among multiple versions of a text, which one was the original for a new modified version.

Keywords: digital humanities, graphs, versioning, editing.

1 INTRODUCTION

With the development of technology and the use of the Internet for not only the retrieval of information, but also for making information available to users, documents that were once only in print are now becoming available for ease of access online, changing the publishing industry distribution channels. Digital Humanities has played a significant role in this change with the traditional areas of humanities utilizing computer technology within their works and studies [6]. Researchers and scholars in the humanities field have now fast access to variations of the same text. Any search conducted for a specific topic may provide several links to view the requested information. A literary critic, searching for a specific work, will more than likely be provided with multiple links. Those links may refer to almost identical texts, sometimes just one or two words different. If those texts are known to have been written by the same author, the possibility of plagiarism should be excluded and the texts be considered versions of one another. The issue becomes to determine which text is the original and which is a version of the original. It is possible that the author has left a no record of the evolutionary path of his or her work, usually known as versioning. If such information does not exist, it can become an issue for scholars performing literary analysis of the author's writing.

Vitali defines versioning as the management of multiple copies of the same evolving online resource, captured at different stages of its evolution. His study is not concerned on who created a text or when exactly it was created but more in the sense of when the information was made available online and when it was accessed and used as a reference point. Access to older versions of an artifact usually allows for accountability and verification of authoring activities. Vitali main goal is to improve web searches, using a time stamp for date of creation or last modification to help the user to determine its relevance. However, to solve the problem of an uncertain time stamp, Vitali uses a formula to calculate a confidence value based on the contents of the document itself [10]. However, his investigation on when the text was made available online does not imply from which previous version it was derived.

In a more academic approach, Buttleer focuses on the changes of the actual structure of a document, how it appears, instead of what the document contains [1]. Document structure can also play an important role in literary analysis. Therefore, Buttleer's research on computational algorithms to determine a minimum edit cost between documents is of significant importance for the literary critic, lacking however, a prediction of its evolutionary sequence.

To keep track of a text evolution, Schmidt and Colomb developed a multi-version document model to represent versions of works. This multi-version document model takes into consideration four operations: insertion, deletion, substitution, and transposition. It was developed to address some of the technical deficiencies of markup languages used within cultural heritage texts [5]. Their work demonstrates the initial use of a graph to represent the reasoning behind a document version.

As an example of the significance of versioning, Vetter and McDonald investigate the work by Emily Dickinson due to her use of implicit meaning and themes to variation of words, phrases, lines and line groups. They use the Text Encoding Initiative as their foundation for encoding the text [8, 9]. Their study demonstrates the use of automated tools to show the complexity of Dickinson's poetic epistolary form.

In another study of the evolution of the texts, Wilson focused on the Regressive Imagery Dictionary, which was written in English to determine the validity of its translations. There had been a wide range of validity studies based on English versions but not in its translations. The translations validity was based on comparative studies of its original with Latin, German and Portuguese versions [11]. While his conclusions are significant to the humanities field, it does not provide any data on the evolutionary path of the translations, which were assumed to be known.

Short and Semino discuss how an author writing style can also be used in the process of text evaluation, based on the comparison of different versions of the same work. In particular, they analyze the poem *The Tyger* by William Blake, by taking two versions that are different by two words [10]. In this case, the evolutionary sequence was already known and their study focused on comparing those texts and analyzing the reasons for the changes applied to the first version.

In a different context, Darwin's Theory of Evolution stresses that the development of life is done by the modification of existing species [2]. In a concise view, it states that mutations occur within an organism's genetic code over time resulting in an different organism, if not just a variation of the original genetic code. Similarly to Darwin's Theory, this study focuses on the evolution of text documents. Text documents have different versions or variations similar to the mutations described by Darwin. Such variations are due to different reasons, such as geographical purposes, social conditions (translation, war, economy), and time progression (modern views). Authors have different reasons as to why modify their original work, but changes that they make to a text can be seen as a mutation or in this study, a version of the document. Essentially, the most recent products are just a modification from the first, still with the same concept, but possible changes in words, punctuation, or writing style. The most important, such changes may have been applied to the original version or to later versions, creating a evolutionary sequence.

In a previous research, conducted in the area of Artificial Intelligence, the authors describe the use of Dempster-Shafer Theory in identifying the original version and the evolutionary sequence of the texts [3]. That process was based on the number of changed characters between two texts and their ratio to the total number of unchanged characters. Those quantities allowed to calculate the mass, belief, plausibility, evidence interval, and doubt, which are significant components on the determination of the possible original document and its evolutionary sequence. The complexity of that solution was high, making its applicability restricted to short texts. This study proposes a solution based on a minimum spanning tree algorithm, reducing such complexity. The next section provides a brief review of the minimum spanning tree algorithm, followed by a section describing the graph model used to represent the text versions. The evolutionary sequence prediction is presented next, followed by a simple example and a summary of the paper.

2 BACKGROUND

Considering a connected graph G = (V, E) where V is the set of vertices and E is the set of edges, a minimum spanning tree (MST) is usually defined as an acyclic subgraph M = (V, T) of G where $T \subseteq E$ such that T connects all vertices and the total weight of its edges is minimum. Two common algorithms used to find a MST are the Kruskal's algorithm and the Prim's algorithm [4]. A simple pseudocode representation of the Kruskal's algorithm is shown in Figure 1.

Figure 2 shows an example of a graph with 5 nodes and 7 edges, which will be reduced to its minimum spanning tree by following the Kruskal's algorithm.

After initializing the forest F with the 5 trees (nodes) and sorting the edges, the first edge selected is AB since it has the least weighted edge. The forest F will now have 4 trees, AB, C, D and E. The next edge to be selected will be AD with a weight of 5, restructuring F to ABD, C and E. Next we connect CE because it contains the least weighted edge of 6, and F will consist of ABD and CE. Lastly, a connection between E and B is selected since the weight 7 is less than BC with 8 and DE with a weight 15. BD is not considered since B and D already belongs to the same tree in F. The result is the minimum spanning tree shown in Figure 3.

```
Kruskal(V, E, w)
$T \leftarrow 0$
create a forest F where each v is a tree
sort E into non-decreasing order by weight w
for each $(u, v) \in E$ taken from the sorted list
        if (u,v) connects 2 different trees in F
                then $T \leftarrow T \cup \{(u, v)\}$
                        $F \leftarrow F \cup \{(u, v)\}$
return MST = (V,T)
```

Figure 1. Kruskal's Algorithm

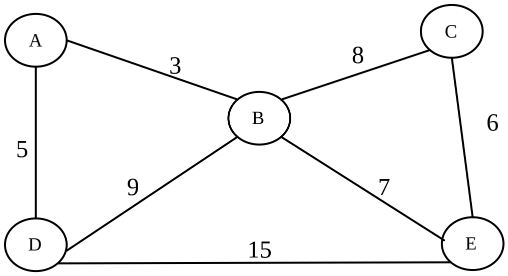

Figure 2. Example graph

3 GRAPH CONSTRUCTION

In order to investigate the evolutionary sequence of the text versions, a text comparison graph (TCG) is constructed to represent the differences between the texts.

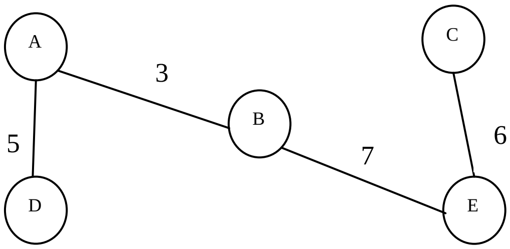

Figure 3. Minimum spanning tree for Figure 2 graph

Definition 1: A text comparison graph (TCG) is an undirected fully connected graph G = (V, E, W) where V represents a set of nodes, representing each of the texts being compared, E is a set of edges where each edge represents the relationship between any two texts, and W is a weight function from E to N that returns the number of modifications between the two texts. In order to compute the values of the function W, one needs to compare the individual characters of two texts, excluded all spaces and space equivalent symbols, such as newlines and tabs, recording the number of insertions and deletions. Substitutions are considered to be the deletion of one character followed by the insertion of a new character.

A simple example considers the texts T_1 and T_2 as follows:
T_1 = "Two dozen sweet bananas."
T_2 = "Twenty four sweet yellow bananas."

The words "Two dozen" in T_1 are replaced by "Twenty four" in T_2, which implies that the characters "odoz" in T_1 were deleted and the characters "tyfour" were inserted in T_2 for a total of four deletions and 6 insertions. The words "sweet" and "bananas" are the same in each text, while "yellow" which contains six characters was added to T_2. Therefore 16 operations were accounted for, resulting in W (T_1 -T_2) = 16. The TCG graph can be seen in Figure 4.

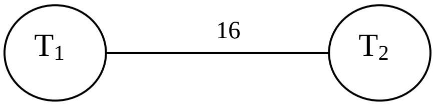

Figure 4. Simple example of TCG graph.

4 EVOLUTIONARY SEQUENCE PREDICTION

In a situation where only two similar texts exist and one of them is known to be the original, it is obvious that the second text is a version of the first. However, when three or more texts are considered, being one of them known to be the original, an assumption has to be made on how to identify the version sequence. Such an assumption is based on the similarity of the texts. The least number of modifications should be indicative of a direct derivation of the original text. Without loss of generality, consider three texts A, B and C, where A is the original text. Assume W(A-B) = n, W(A-C) = m, and

W(B-C) = r. If n < m then a conclusion can be made that B is a version of A. However, the decision about C, whether a version of A or B, depends on the value of r. If r < m then C is a version of B, otherwise C is also a version of A.

Lemma 1: Given a TCG G= (V, E) where V={T_1, T_2} such that W(T_1-T_2) = 0 then the texts are identical.

Such Lemma can be easily proven by contradiction.

Lemma 2: Given a TCG G= (V, E) where V={T_1, T_2, T_3} such that W(T_1-T_2) = n, W(T_1-T_3) = m, W(T_2-T_3) = r, as shown in Figure 5, where T_1 is the original text, if n < m, and n < r then T_2 is a version of T_1.

Given the similarities between T_1 and T_2, lower number of insertions and deletions between them, Lemma 2 is also easily proven by contradiction.

Lemma 3: Given a TCG G= (V, E) where V={T_1, T_2, T_3} such that W(T_1-T_2) = n, W(T_1-T_3) = m, W(T_2-T_3) = r, where T_1 is the original text and T_2 is a version of T_1, if T_3 is a version of T_2 then r < m.

The proof for this lemma can be derived from the fact that there are two possible evolutionary paths from T_1 to T_3. A direct version from T_1 would require m changes in the text, while a path through T_2 requires n + r changes to T_1. From these two paths one can assume that m replays the n changes from T_1 to T_2, added to the r changes from T_2 to T_3 making m greater than r.

Considering Lemmas 1, 2, and 3, it is easily shown that the lower weight edges of a TCG are the expected indications of the evolutionary versioning, justifying the use of the Kruskal’s algorithm in determining the minimum spanning tree and consequently the predicted version path.

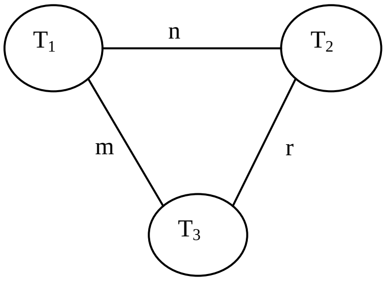

Figure 5. Lemma 2 Graph

5 EXAMPLE

A simple example involves five versions of the poem “Faith is a fine invention,” written by Emily Dickinson around 1860 and originally published in 1891 after her death. On a simple online search, different versions of this poem can be found on different websites. The five different texts below were retrieved on November 2011.

From http://www.poemhunter.com/poem/faith-is-a-fine-invention/, designated poem 1:

"Faith" is a fine invention
When Gentlemen can see—
But Microscopes are prudent
In an Emergency.

From http://www.angelfire.com/ca/kesheng/faith.html designated poem 2:

“Faith" is a fine invention
when Gentlemen can see --
But Microscopes are prudent
In an Emergency.

From http://www.goodreads.com/quotes/show/44777 designated poem 3:

“Faith is a fine invention
When gentlemen can see,
But microscopes are prudent
In an emergency.”

From http://www.online-literature.com/dickinson/ poems-series-2/32/ designated poem 4:

Faith is a fine invention
For gentlemen who see;
But microscopes are prudent
In an emergency!

From http://www.emilydickinsonmuseum.org/church designated poem 5:

"Faith" is a fine invention
For Gentlemen who see!
But Microscopes are prudent
In an Emergency!

These five poems can be represented as nodes of a fully connected graph, as seen in Figure 6. The weights were calculated based on the number of insertions and deletions required to modify one text and produce the next text. Poem 1 is known to be the original text.

From the minimum spanning tree it can be seen that Poems 2, 3 and 5 are expected versions of Poem 1 and Poem 4 is an expected version of Poem 5.

6 SUMMARY

In the analysis of literary works from the same author it is important to know the evolutionary sequence, versioning sequence, of the texts. In many situations, such sequence is unknown. This study presented the use of a minimum spanning tree algorithm in determining a possible version sequence. Such solution requires the modeling of a text comparison relation to a graph whose edges are labeled with weights measuring the number of inclusions and deletions required to modify the text in order to create a similar text. Literary critics can use such solution as the basis to evaluate the reasons why an author created a new version of his or her work.

7 REFERENCES

[1] Buttler, D. "A Short Survey of Document Structure Similarity Algorithms," The 5th International Conference on Internet Computing, Las Vegas, NV, June 2004.

[2] Greenberger, R. "Darwin And The Theory Of Evolution," The Rosen Publishing Group, Inc., New York, 2005.

[3] Mayfield, K. Grant, E., Albert, C., Kim, E. "Identifying the Evolution Sequence of a Text Document" In the Proceedings of the 5th International Multi-Conference on Engineering and Technological Innovation, pp. 184-189, Florida, July 2012.

[4] Neapolitan, R., Naimipour, K. "Foundations of Algorithms Using C++ Pseudocode," Jones and Bartlett, Inc., New York, 2004.

[5] Schmidt, D., and Colomb, R. "A data structure for representing multi-version texts online," IJHCS '09, pp. 497-514

[6] Schreibman, S., Siemens, R., Unsworth, J. "A companion to Digital Humanities," Oxford: Blackwell, 2004.

[7] Short, M., Semino, E. "Evaluation and stylistic analysis," The Quality of Literature: Linguistic Studies in Literary Evaluation, John Benjamins Publishing Co., pp. 117-137, May 2008.

[8] Text Encoding Initiative. http://www.tei-c.org/ index.xml. June 29, 2011.

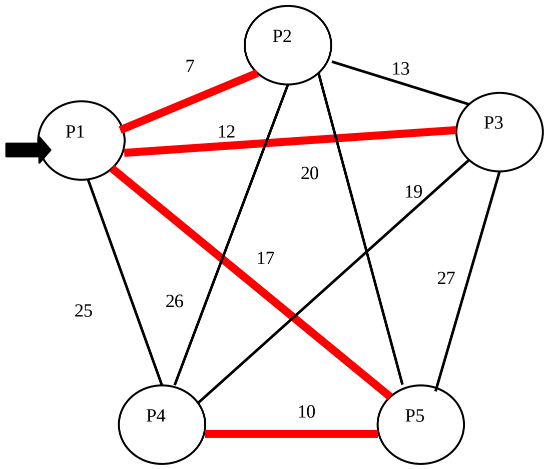

Figure 6. Five poem graph representation and MST

[9] Vetter, L., and McDonald, J. "Witnessing Dickinson's Witnesses," Literary and Linguistic Computing, Vol. 18, No 2, pp. 151-165, 2003.

[10] Vitali, F. "Versioning hypermedia," ACM Computer Surveys, Vol. 31, pp. 1-7, December 2009.

[11] Wilson, A. "The Regressive Imagery Dictionary: A test of its concurrent validity in English, German, Latin, and Portuguese," Literary and Linguistic Computing, Vol. 26, No. 1, pp. 125-135, 2011.