Identifying the Evolution Sequence of a Text Document

Katia Mayfield Emanuel S. Grant Eunjin Kim
Department of Computer Science, University of North Dakota
Grand Forks, ND 58201

and

Crystal Alberts
Department of English, University of North Dakota
Grand Forks, ND 58201

ABSTRACT

An issue that has been around in the case of published documents and now with the technology of the Internet is the management of versions of data being created or digitized. The field of Digital Humanities concentrates on the digitization of documents and also supports the creation of "born digital" documents. Considering the texts being digitized and those being created electronically, it can become difficult to determine which published (whether on paper or digital) document is an original, which one is a version of the original, and which one may be a version of a version. In this paper, a sample case of the Emily Dickinson poem *Faith is a fine invention* is analyzed by using the Dempster-Shafer's theory in order to identify the evolution of three versions of the poem.

Keywords: Dempster-Shafer's theory, Digital Humanities, born digital, Text Encoding Initiative, Electronic Literature Directory.

1. INTRODUCTION

One of the main areas of study for scholars in the area of English, under the field of Humanities, is the study of authors and their work. There are several literary works that have different versions created by their own authors or by other agents such as publishers, translators and editors. Depending on the historical time frame of these works, today's scholars may have a difficult time trying to distinguish which one may be the original and which versions were created from the original or from other previous versions. With the new area of Digital Humanities, many documents are now becoming digitized and this creates better chances for scholars to be able to conduct their research by accessing documents electronically, rather than having to travel to locations that house different historical pieces. This electronic access allows the identification of multiple versions of the same text. However, it does not establish the evolutionary path of the document.

Digital Humanities, also sometimes referred to as Digital Computing, is a new field of study where the traditional area of Humanities – English, history, anthropology, literature, etc – work in collaboration with computational sciences [9]. The purpose of this field is to study and implement ways to archive documents that are found on paper or those that are "born digital". "Born digital" data are those that are created by using multimedia systems, basically, that are created on a computer and are specifically made to only be viewable on a computer [9].

According to Matthew Kirschenbaum, within the Humanities field, English departments have been the one's to take advantage of concepts such as Digital Humanities. He gives five reasons to support such claim: text is the most tractable data type to be manipulated by computers, the association between computers and composition, the convergence between editorial theory and method from the 1980s and the means to implement electronic archives and editions of documents, projects around hypertext and other forms of electronic literature, and last is an openness of English departments to be involved in cultural studies where computers and other digital material become centerpieces of analysis [5].

One of the topics that English scholars may focus on is the study of an authors' life. In this area, they may find that an author may have written a piece of work (whether digital or not) and that a few years later the author may have re-written that same piece, creating a different version to it. One well known author that has done this with several of her poems was Elizabeth Barrett Browning. A sample of her poems that can be found to have more than one version are *To Flush, My Dog* (1843, 1844, 1850, 1853, and 1856), *A Child*

Asleep (1840, 1844, 1850, and 1853), *Loved Once* (1844, 1850, and 1853), *To Bettine* (1850, and 1853) [11]. The identification of such different renderings of the same text has been the focus of several studies in Digital Humanities.

Studying how to handle these different versions, Dyreson, Lin and Wang introduced a computational tool that tracks changes to documents, recording modifications applied to the original document or to a previous version [4]. While the tool seems to be very efficient, it requires the versions to be developed under the tool in order to be effective. In another project in this area, Anick and Flynn discuss the implementation of data structures designed to improve the query of database systems where the data content may have been modified, creating different versions of such content [1]. As in the tool proposed by Dyreson, the versioning of the data requires time stamps to be recorded during the modification of the contents [4].

Other studies, covering documents that could have more than one version, did not report such occurrences. For example, Antonacopoulos et. al published a study on the procedures used to transform historical written documents on electronic ones [2]. The study focuses on the recovery of the scanned original text and does not mention the possibility of the existence of more than one version of the same document. Also, Crane, Smith and Wulman describe the implementation of a digital humanities project, involving books, images and maps of London [3]. Such work demonstrates the importance of the digitization of historical documents. However, the authors do not mention the existence of duplicates or second versions of the digitized documents.

This paper focuses on sets of versions of the same text, discussing the use of uncertainty reasoning to identify original documents and the sequence in which versions evolved from that original. A brief introduction to Digital Humanities and a short description of the Dempster-Shafer Theory of Evidence and its association to Belief functions, used in uncertainty reasoning, are presented in the next section. They are followed by the presentation of a poem by Emily Dickinson, which will be used as a case study in this paper. A description on how Belief functions were used to determine the original work and the sequence in which it was modified follows. Finally, a summary of the presented material closes the study.

2. BACKGROUND

According to Kirschenbaum, Digital Humanities is about "scholarship (and a pedagogy) that is publicly visible in ways to which are generally unaccustomed, a scholarship and pedagogy that are bound up with infrastructure in ways that are deeper and more explicit than what scholars are typically accustomed to, a scholarship and pedagogy that are collaborative and depend on networks of people and that live an active 24/7 life online" [5]. Even with this definition of modern digital humanities, there were two challenges that had to be met, those were the establishment of standards and storage.

Today, documents are being created as "born digital" data and old documents are being converted to electronic format for the purposes of archiving. One of the main issues with it all becoming available electronically is establishing a standard that authors and those converting documents should follow. Just like programming with a particular computer language requires a syntax that must be followed for the program to work correctly, with Digital Humanities some authors choose to encode their work with Extensible Markup Language (XML), which has its similarities to HyperText Markup Language (HTML). In particular, when HTML is used, there are standards that must be followed and creators can validate their code through the World Wide Web Consortium [12]. While both of these languages are very similar, XML is used to 'tag' information to make it searchable online while HTML tags provide tools for a web designer to specify how a web page should render. However, XML has no current universal validator like that of W3C for HTML. Considering this being an issue, the Text Encoding Initiative was created to try to establish some sort of standard that those coding in XML should be able to use and "validate" their code [10]. Even though TEI is not a standard used universally, those in the Humanities field are familiar with it and follow these standards when encoding their documents.

Besides trying to establish standards to be followed in this new area, there is also a reasonable concern about storage. There are funded projects that are working on this issue, and there are organizations that have been established to help. The Electronic Literature Directory was created to provide a database for these digital documents to be housed [8]. To submit their documents to the ELD, creators will identify the techniques that were used to create their document, and are required to use XML encoding in their documents where appropriate.

The 24/7 ease of access that Digital Humanities has created for "born digital" documents and those that have been archived, makes it easier not just for scholarly work to be done but also for duplication of

work. Here, it is not being referred to copyright infringements taking place, but by duplication of work considering that there is ease of access, allowing for those who have permissions to work on a document, to be able to collaborate, or work on their own, to create or identify different versions of the document. In such situation, scholars, being able to find different versions of texts that they are researching, may bring up the question of which document is the original, which is a version created from the original, and which are versions created from other version. Scholars may end up with a sense of doubt and uncertainty, which then requires them to conduct additional research to establish the appropriate order of documentation for their research to be accurate. Due to possible lack of information, such order may not be considered exact.

When taking the concept of uncertainty into consideration, in the field of artificial intelligence there are three types of uncertainty that should be considered. First is nonspecificity (or imprecision), which is connected with sized (cardinalities) of relevant sets of alternatives; next is fuzziness (or vagueness), which results from imprecise boundaries of fuzzy sets; and lastly is strife (or discord), which expresses conflicts among the various sets of alternatives [6]. This last type of uncertainty, strife, is the one which best describes the uncertainty being considered among the different versions of documents being examined by scholars.

For the purpose of this paper, the concept of uncertainty is associated to the Dempster-Shafer's theory (DST) which is a mathematical theory of evidence [6]. This theory is based on two dual nonadditive measures: belief measures and plausibility measures [6]. Shafer's framework allows for belief about propositions to be represented as intervals, bounded by two values, *belief* (or *support*) and *plausibility*:

$$belief \leq plausibility$$

The basis for understanding belief and plausibility is to think of having a hypothesis that one must work the basis of their scholarly work on. If we define mass as the proportion of available evidence that supports a claim, the belief measurement is the sum of masses, which are held by all the subsets included in the hypothesis. The belief of a hypothesis will form a lower bound, being its amount that directly supports the hypothesis. The plausibility is considered to be an upper bound of the possibility that the hypothesis is true. By calculating degrees of belief and plausibility on items that describe the sequencing of different versions, a scholar should be able to come to a satisfactory conclusion of the order in which documents were created.

The formal definitions of the Dempster-Shafer's theory establish several mathematical concepts. If one considers X to be universal set: the set representing all possible cases of a problem solution then power set 2^x, known as the power set of X, is the set of all subsets of X, including the empty set. In the Dempster-Schafer theory, each element in the power set can represent the scheme of the state of the system, by representing the states in which the proposition is true. The theory of evidence requires for an expert that can analyze the "scenario" or the existence of statistical data. Based on the expert report or the statistical data, each element that is included in the power set is given a belief mass, which is a value found in the range of [0, 1], and this can be represented by a function:

$$m: 2^x \rightarrow [0, 1]$$

This function is the basic belief assignment and it has a mass for the empty set which is 0, and a second mass for the remaining members of the power set, which all sum up to 1.

$$\sum_{A \in 2^x} m(A) = 1$$

The Dempster-Shafer's theory makes the claim that whichever states belong to the set A but to no specific subset of A. The mass m(A) expresses the proportion of evidence that supports this claim. By assigning masses, the probability interval can be given an upper and lower bound that it falls between. The upper bound is created by the belief measure and the lower bound by the plausibility measure.

$$bel(A) \leq P(A) \leq pl(A)$$

The belief bel(A) for a set A is defined as the sum of all the masses of subsets of the set of interest:

$$\sum_{B|B \subseteq A} m(B)$$

The plausibility pl(A) is the sum of all the masses of the sets B that intersect the set of interest A:

$$\sum_{B|B \cap A \neq \emptyset} m(B),$$

The two measures are related to each other as follows:

$$pl(A) = 1 - bel(\bar{A})$$

To help with analysis, experts can also take into consideration the Evidence Interval. This interval uses the belief and plausibility as minimum and maximum bounds respective. Between these values is where the probability is shown to be true.

$$EI(A) = [bel(A), pl(A)] = [bel(A), 1 - bel(\overline{(A)})]$$

The degree to which A can be disbelieved or refuted can also be calculated and is referred to as Doubt or Dubiety:

$$Dbt(A) = bel(\overline{(A)}) = 1- pl(A)$$

The application of these formulas and theories to a sample case of versioning is discussed in Sections 3 and 4 of this paper.

3. SAMPLE CASE

When documents have more than one version to it, it may become difficult to be able to distinguish which version was created from the original and which versions may have been created from another version. This problem can be considered one of doubt or better yet "uncertainty". According to Klir, "uncertainty can be caused by information being incomplete, imprecise, fragmentary, not fully reliable, vague, contradictory, or deficient in some other way" [7]. Something that can be done is to find a way to be able to measure the amount of uncertainty, focusing next on a new goal which would be to be able to reduce this value of uncertainty.

The poem "Faith is a fine invention" written by Emily Dickinson circa 1860 and originally published in 1891, after her death, is used to demonstrate the concept of versioning.

On a simple online search, different versions of this poem can be extracted from several websites and the different forms it assumes can be easily identified. Three of these different texts, retrieved on November 2011, are:

From http://www.goodreads.com/quotes/show/44777 designated poem 1:

"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 2:

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

From http://www.emilydickinsonmuseum.org/church designated poem 3:

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

Such poems were target of this study in order to show how the proposed belief functions work in the identification of the versioning process. There are two problems associated to these different versions. The first is to identify the original version, the second is to try to identify which ones were modified to produce the consecutive versions. Therefore, the solution to be considered is that one of the poems is the original in this evolution. Then, assuming that the notation x → y indicates that y is a version derived from x, a relation may be defined showing all possible pairings. However, for a number n of poems, the number of different combination (pairs) of version and original is $n!/(n-2)!$ or simply $n*(n-1)$. In a simple case, for 5 poems, there would be 20 combinations. Then, the superset, which is required in the Dempster-Shafer theory, would have $2^{n(n-1)}$ (in the example $2^{20} \approx$ 1,000,000) elements.

Due to the size of possible combinations to be examined on the Emily Dickinson example, this study was limited to three versions of the poem (designated from now on poem 1, poem 2 and poem 3). In such situation, there are 64 powerset combinations that need to be considered.

In the case of the three poems, six variables were defined to represent the combination pairs:

A = P1→ P2
B = P1→P3
C = P2→P3
D = P2→ P1
E = P3→P1
F = P3→P2

According to the Dempster-Shafer theory, each element of the superset must have a mass value associated to it. In this study, an evidential value was calculated by counting the number of unchanged characters of the poem assumed to be the version (y), based on the poem assumed to be the original (x).

In order to identify the sequence of versions a two-dimensional matrix was created to record the number of unchanged characters between every pair of poems, as seen in Table 1.

Table 1. Unchanged characters between poems

Versions:	Poem 1	Poem 2	Poem 3
Original			
Poem 1	-	83	80
Poem 2	82	-	87
Poem 3	79	87	-

Once the values of the variables A to F were calculated, directly from Table 1, the combinations of those variables had their evidential value computed by adding the individual values. Then, the mass of each variable was established by computing the fraction between the unchanged characters and the total number of evidential values, creating a fraction in the range 0 to 1, with the property of having the summation of all masses equals to 1. Basically, the matrix entries were used to establish the mass (m numbers) to support the beliefs that two versions are connected someway, using the formula

$$m(Z)= \frac{matrix\ count(x,y)}{summation\ value\ of\ matrix\ count},$$

which gives a stronger value to those with smaller changes.

The belief and plausibility of every variable and set could then be calculated using the formulas presented in Section 2. Table 2 shows some of those values.

4. ANALYSIS

In this section, the Dempster-Shafer theory is applied to identify which of the poems is the original version. This is a hypothetical situation where the data was artificially established based on a simplistic assumption that the number of unchanged characters would be less when two poems were related by versioning. This simple assumption allows a demonstration of the effectiveness of the proposed method.

Analyzing the results of the different calculations, the obvious conclusion is that the correct result will come from the variables representing versioning pairs, under the condition that a poem should be the version of another poem and the chosen answer could not contradict itself.

Based on the highest belief value and lowest doubt, it seems that the combinations found in F and C would represent the correct chronological order. However, these two are inverses of each other, and therefore, a selection of one of them must imply the elimination of the other. To further analyze the information and determine a better order, both values are evaluated.

When F is chosen, it represents that poem 4 is a version of poem 5. C at this point is eliminated because it is the inverse of the selected variable F, and A would also be eliminated here because poem 4 cannot have two origins. The next variable with the highest belief is D where poem 3 is a version of poem 4, and the last option to be looked at is E, which shows poem 3 as a version of poem 5. By taking into comparison the values calculated for D and E, it is seen that D has a higher belief, plausibility, evidence interval and lower doubt. Therefore by having a higher belief value, and a lower doubt value, the logical conclusion would be to select D, and therefore the order would be that poem 5 is the original to poem 4 and poem 4 is the original to poem 3.

Next we analyze the information based on variable C, in this case variable F would be eliminated as it is the inverse of C. A would be the next available variable with the highest belief, and therefore D would be eliminated for being its inverse.

Table 2: Partial table of calculated values

Combinations	Unchanged characters (size 93)	changed characters	Mass	Belief	Plausibility	Evidence Interval	Doubt
F	87	6	0.005459	0.005459	0.61314	0.607681	0.38686
C	87	6	0.005459	0.005459	0.61314	0.607681	0.38686
A	83	10	0.005208	0.005208	0.609375	0.603916	0.390625
D	82	11	0.005146	0.005146	0.608434	0.602974	0.391566
B	80	13	0.00502	0.00502	0.606551	0.601092	0.393449
E	79	14	0.004957	0.004957	0.60561	0.600151	0.39439

In this case B would also be eliminated because one poem cannot have two origins, and by selecting B, poem 5 would be evaluated as originating from both, poem 3 and poem 4, which is not possible. In this case, E would be a second possible variable to consider. By comparing A and E, it is seen that A has higher values in all aspects of calculations compared to E. However, there are two values that negatively affect the variable A (evidence interval being higher, and doubt being higher than the values found for E) so it is excluded and variable E is evaluated. In this case it is shown that poem 5 is a version of poem 4 and poem 3 is a version of poem 5.

The conclusions based on these evaluations are contradicting, except for the fact that without prejudice it can be stated that poem 3 is **not** an original poem to either of the other two poems. The difficult part at this point is to determine if it is the first chronological (FD) order or the second (CE) that will result in a feasible statement. It is seen that the values for F and C are exactly the same for all of the calculations and therefore there is no argument that either one of the two could be chosen. However, the comparison must be made according to the variables of D and E, which are the ones that have different values. This comparison has already been shown that D has a higher belief value than E and also has a lower doubt value than E. As a result, it is concluded that the first analysis, with F, is the one that holds true and the proper chronological order of the poems is poem 5 being the original, poem 4 derived from poem 5, and poem 3 derived from poem 4.

Finally, we call the reader's attention to the fact that his research is limited to only one expert and one way of evaluating the poems. A reliable system should depend in more than one set of data and therefore a second, or more, experts, which would allow the analyzer to calculate combined values and be able to obtain a more definite conclusion.

5. SUMMARY

With traditional ways of publishing documents, authors always had the opportunity to decide to re-publish their work with minor changes to them. At the same time, with new technology, those traditional works can now be made available electronically, while authors have the option to create new texts exclusively in electronic format. Frequently, scholars, studying authors and their works, do not have the opportunity to contact the authors to find the chronological order of their texts or the process taken to create new versions of the document. While at times publication dates are associated with the data, it is not a definite conclusion that a later publication is a version of the original, it just means that it became available at a later date. Since scholars do not usually have any other source besides the texts, they depend on techniques that can determine the evolution of a document. This study shows that with the concept of artificial intelligence and the Dempster-Shafer theory, scholarly experts that analyze versions of texts should be able to obtain a reasonable evolution series for documents with different versions.

6. REFERENCES

[1] Anick, P., Flynn, R. "Versioning a Full-text Information Retrieval System," 15th Annual International SIGIR, pp. 98 – 111, Denmark, 1992.

[2] Antonacopoulos , A., Karatzas, D., Krawczyk, H., and Wiszniewski, B. "The Lifecycle of a Digital Historical Document: Structure and Content," DocEng '04, pp. 147-154, 2004.

[3] Crane, G., Smith, D., Wulfman, C. "Building a Hypertextual Digital Library in the Humanities: A Case Study on London." JCDL '01, pp. 426-434, Roanoke, Virginia, 2001.

[4] Dyreson, C., Lin, H., Wang, Y. "Managing Versions of Web Documents in a Transaction-time Web Server," New York, pp. 422-432, May 17-22, 2004.

[5] Kirschenbaum, M. "What is Digital Humanities and What's It Doing in English Departments?," ADE Bulletin, 2010.

[6] Klir, G. "Uncertainty and Information: Foundations of Generalized Information Theory," Wiley – IEEE Press, 2005.

[7] Klir, G., and Yuan, B. "Fuzzy Sets and Fuzzy Logic: Theory and Applications," Prentice Hall, 1995.

[8] National Endowment for the Humanities, Electronic Literature Directory. http://directory. eliterature.org/. June 29, 2011.

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

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

[11] The University of North Dakota: Elizabeth Barrett Browning Project. http://und.edu/instruct/ sdonaldson/ November 15, 2011.

[12]World Wide Web Consortium (W3C). http://www.w3c.org. November 15, 2011.