Visual Analysis of Conflicting Opinions

Chaomei Chen[1] Fidelia Ibekwe-SanJuan[2] Eric SanJuan[3] Chris Weaver[4]

Drexel University, USA Université de Lyon 3, France Université d'Avignon, France Penn State University, USA

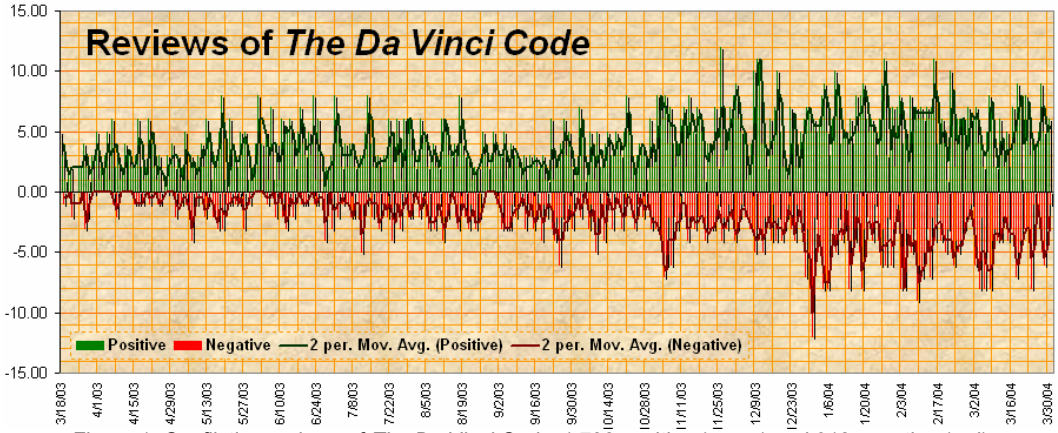

Figure 1: Conflicting reviews of *The Da Vinci Code*: 1,738 positive (green) and 918 negative (red).

ABSTRACT

Understanding the nature and dynamics of conflicting opinions is a profound and challenging issue. In this paper we address several aspects of the issue through a study of more than 3,000 Amazon customer reviews of the controversial bestseller *The Da Vinci Code,* including 1,738 positive and 918 negative reviews. The study is motivated by critical questions such as: What are the differences between positive and negative reviews? What is the origin of a particular opinion? How do these opinions change over time? To what extent can differentiating features be identified from unstructured text? How accurately can these features predict the category of a review? We first analyze terminology variations in these reviews in terms of syntactic, semantic, and statistic associations identified by TermWatch and use term variation patterns to depict underlying topics. We then select the most predictive terms based on log likelihood tests and demonstrate that this small set of terms classifies over 70% of the conflicting reviews correctly. This feature selection process reduces the dimensionality of the feature space from more than 20,000 dimensions to a couple of hundreds. We utilize automatically generated decision trees to facilitate the understanding of conflicting opinions in terms of these highly predictive terms. This study also uses a number of visualization and modeling tools to identify not only what positive and negative reviews have in common, but also they differ and evolve over time.

CR Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Linguistic Processing; H.5.2 [Information Interfaces and Presentation]: User Interfaces —Graphical User Interfaces.

Keywords: Visual analytics, conflicting opinions, terminology variation, decision tree, predictive text analysis, sense making.

1. e-mail: chaomei.chen@cis.drexel.edu
2. e-mail: ibekwe@univ-lyon3.fr
3. e-mail: eric.sanjuan@lia-univ-avignon.fr
4. e-mail: cew15@psu.edu

IEEE Symposium on Visual Analytics Science and Technology 2006
October 31 - November 2, Baltimore, MD, USA
1-4244-0592-0/06/$20.00 © 2006 IEEE

1 INTRODUCTION

Understanding the nature and dynamics of conflicting opinions is a far-reaching and challenging issue. Contradictory opinions exist in a diverse range of domains, including science, engineering, sociology, politics, and international relations. Critical challenges are identifying the basic premises of arguments, assessing the credibility of available evidence, understanding the context and background of a particular position, and tracking the development of how various opinions interact with a broader context of information over time. The ability to accomplish these tasks effectively and efficiently has a direct impact on people's understanding, interpretation, and decision making activities. While detecting trends and dynamics of change attracts an increasing interest, fundamental challenges remain at both macroscopic and microscopic levels due to the dynamic and complex nature of our perception and cognition. Recently, visual analytics has rapidly evolved to meet the need for national security, emergency and disaster preparedness and response, and more traditional science and technology indicators, paradigm shifts in scientific knowledge domains [1].

Reviews of a controversial bestseller book such as *The Da Vinci Code* carry the hallmark of conflicting opinions. Reviews made by readers from a diverse range of perspectives provide a valuable source of insight in terms of how people form their opinions and what influences their opinions.

Figure 1 depicts the distribution of positive and negative reviews of *The Da Vinci Code* on Amazon.com between March 18, 2003 and March 30, 2004, the first year of the publication of the book. Although it is obvious that positive reviews consistently outnumbered negative ones, arguments and reasons behind these reviews are not apparent.

Understanding conflicting book reviews has implications far beyond books, ranging from opinions on merchandise, electronic devices, information services, to opinions on wars, religious, and environmental issues. Advances in this area have the potential of making substantial contributions to the assessment of the underlying credibility of evidence, the strength of arguments, diverse perspectives, and expectations.

In this paper, we present a study of positive and negative reviews in order to identify technical challenges and improve our understanding of contradicting opinions. Choosing this topic has distinct advantages: no prior domain knowledge required, easy to interpret and evaluate results, potentially extensible applications to other genres. The rest of the paper is organized as follows. We first introduce related work and existing approaches. Then we describe key concepts and major components of our approach, followed by methods and results of this study. Finally, we identify the strengths and weaknesses of the current approach.

2 Related Work

Several areas of research are relevant, in particular, text mining, information diffusion and tipping points, sentimental analysis of movie reviews, sense-making and visual exploration [2].

The role of invisible colleges in the diffusion of scientific knowledge has been the subject of a variety of research. For example, co-citation analysis studies thematic changes in scientific literature by relying on citation links as an indexing mechanism [3]. Researchers in the text mining and information exploration communities have also studied ways to identify influential papers without the presence of references [4]. A well-known system for visual information exploration is IN-SPIRE[1]. One of the fundamental challenges is making sense of a high dimensional thematic space and dealing with the complexity that exceeds human perceptual and cognitive capabilities. For instance, Latent Semantic Analysis (LSA) is a powerful dimensionality reduction method. However, empirical evidence shows that it requires as high as 300 latent concepts for LSA to correlate with human judgments of similarities between texts. Understanding the structure of a 300-plus dimensional abstraction remains a challenge to text analysis as well as visualization [5].

Detecting emergent patterns in an open and information rich environment becomes increasingly important as the speed of information diffusion increases. A technical challenge is how to piece together fragmented information and form a big picture. A recent example in this area is BlogPulse [6], which aims to discover trends in weblog entries. BlogPluse relies on the extraction of key phrases, person names, and key paragraphs from weblog texts. BlogPluse identifies a key phrase if the phrase occurs more frequently on a day than its average frequency over the past two weeks.

Understanding the thematic evolution in texts has been studied from several perspectives. ThemeRiver [7] depicts thematic flows over time in a collection of articles. The thematic changes are shown along a time line of corresponding external events. A thematic river consists of frequency streams of terms; the changing width of a stream over time indicates the changes of term occurrences. The occurrence of an external event may be followed by sudden changes of thematic strengths. Kaban and Girolami [8] introduced a probabilistic method based on latent variable models for unsupervised topographic visualization of evolving textual information. Their method can be seen as complementary tool for topic detection and tracking. They applied their method to the study of a chat-line discussion data set. The data is produced in Internet relay chat rooms.

Sentiment analysis is a closely related topic, which aims to identify underlying viewpoints based on sentimental expressions in texts. Pang and Lee [9] presented a good example of classifying movie reviews based on sentiment expressions. Pang and Lee used text-categorization techniques to identify sentimental orientations in a movie review and formulated the problem as finding minimum cuts in graphs. In contrast to previous document-level polarity classification, their approach focuses on context and sentence-level subjectivity detection. The central idea is to determine whether two sentences are coherent in terms of subjectivity. It is also possible to locate key sentimental sentences in movie reviews based on strongly indicative adjectives, such as *outstanding* for a positive review or *terrible* for a negative review. However, such heuristics should be used with considerable caution because there is a danger of overemphasizing the surface value of such cues out of context.

A more recent study [3] describes the design and use of a system to support humanities scholars in their interpretation of literary work. The system integrates text mining, a graphical user interface and visualization. Users can interactively read and rate documents found in a digital libraries collection, prepare training sets, review results of classification algorithms and explore possible indicators and explanations. Their approach is complement to ours in terms of user tasks supported. In this study, we focus on automated approaches to capture essential lines of arguments or debates from a body of unstructured text without human intervention.

The majority of relevant research is built on the assumption that desirable patterns are prominent. Although this is a reasonable assumption for patterns associated with mainstream themes, there are situations in which such assumptions are not viable, for example, detecting rare and even one-time events and differentiating opinions based on their merits rather than the volume of voice.

In this article, we introduce a complementary approach to aid visual analysis of conflicting opinions. Specifically, we build on research in terminology variation and combines with predictive text analysis and interactive visualization techniques to support the understanding, interpretation, and verification of conflicting information from a diverse range of perspectives. We apply this integrative approach to the analysis of customer reviews of *The Da Vinci Code*. And we expect to identify what differentiate positive and negative reviews of the book and temporal dynamics of the development of various themes.

3 Terminology Variation

Terminology variation is a key issue in computational terminology [10]. It focuses on symbolic relations between terms and how they can be related through several types of variations and transformations [11].

3.1 Linguistic Operations

Term variation refers to the transformation of a term to a conceptually related term through linguistic operations such as morphological, syntactic, and semantic operations (See Table 1). Identifying semantic variants requires extra sources of information of semantics, such as, WordNet [12]. Selecting variation relations is a term filtering process because terms are selected only if their variants of some types can be found in the corpus, including co-occurrence variants. Identifying term variations enables us to capture the actual state of knowledge in a given domain. This in turn promotes in-depth and microscopic knowledge discovery and knowledge evolution study.

[1] http://in-spire.pnl.gov/

Table 1: Linguistic operations underlying term variations.

Operations	Term	Term Variation
Morphological (spelling)	page-turning suspense	page turning suspense
Syntactic (adding a modifier)	secret society	ancient secret society
Syntactic (adding a head word)	clever plot	clever plot twist
Syntactic (changing a modifier)	renowned Harvard professor	famous Harvard professor
Syntactic (changing a head word)	secret book	secret agenda
Semantic (synonymous)	ingenious plot	clever plot

3.2 TermWatch

The TermWatch system is originally designed to depict topics from scientific and technological literature [13, 14]. It combines linguistic analysis with a scalable clustering algorithm in order to visualize important topics contained in a text corpus. This symbolic lexico-syntactic approach is particularly suitable for clustering multi-word terms, which rarely re-occur in text. Such terms often lead to very large and sparse matrices that are difficult to handle by existing statistical approaches to clustering which rely on high frequency information.

TermWatch comprises three components: a term extractor, a relation identifier, and a clustering module. All the data are stored in a MySQL database. Term extraction in TermWatch utilizes LTPOS[2]. Then, terminological variations between terms are identified. These identified term variants are subsequently clustered with a hierarchical clustering algorithm called *Classification by Preferential Clustered Link* (CPCL).

The CPCL algorithm operates in two stages. First, conceptually related terms are grouped together. These terms share a common head word, but have different modifiers as in terms *ingenious plot* and *clever plot*. Such groups of terms are linguistically related through a subset of the term variation types called COMP. These components typically include spelling variants, WordNet semantic variants, and modifier variations.

In a second stage, these components are iteratively clustered based on the second subset of term variation types called CLAS. CLAS relations typically indicate a considerable change from one term to another, for example, the change of a head word from *secrete **book*** to *secrete **agenda***. The clustering is based on the number of variations across the components and the frequency of the variation type.

CPCL avoids the well-known chain-effect drawback of single link clustering without losing its intuitiveness and computer tractability. It has been shown that this variant of hierarchical clustering preserves its main ultrametric properties [15]. The clustering algorithm is implemented using a straightforward $\boldsymbol{O}(E)$ procedure called *Select Local Maximum Edge* (SLME) [16].

TermWatch supports optional associations based on term co-occurrence. Co-occurrence links can be combined with any subset of the variation relations. This makes TermWatch a comprehensive platform that combines statistical and symbolic criteria for text data analysis at the microscopic level. Clustering results can be accessed either via an integrated visualization package aiSee[3] for domain topic mapping or through an interactive hypertext interface.

2 http://www.cogsci.ed.ac.uk/~mikheev/tagger_demo.html

3 http://www.aisee.com.

The properties of term variation clusters generated by TermWatch have influenced the choice of a visualization tool. Foremost is the fact that it generates undirected graphs whose layout is determined from the strength of external links between clusters. Since connected components are clustered instead of individual terms, it is necessary to be able to unfold a cluster and reveal its terms. The aiSee visualization package was integrated to the system. The output of the clustering module in TermWatch is automatically formatted in the *Graph Description Language* (GDL) for aiSee visualization. Each cluster can be unfolded to reveal its internal structure in terms of its connected components and the most active variants. The user can explore and examine the most salient features of a cluster.

4 A Visual Analytic Approach

We introduce a visual analytic approach to analyzing conflicting opinions and their temporal dynamics. In particular, we demonstrate the application of this approach to a study of positive and negative customer reviews of *The Da Vinci Code*.

The procedure consists of several steps: data collection, term variation analysis, time series visualization of term variants, classification based on selected terms, and content analysis (See Figure 2).

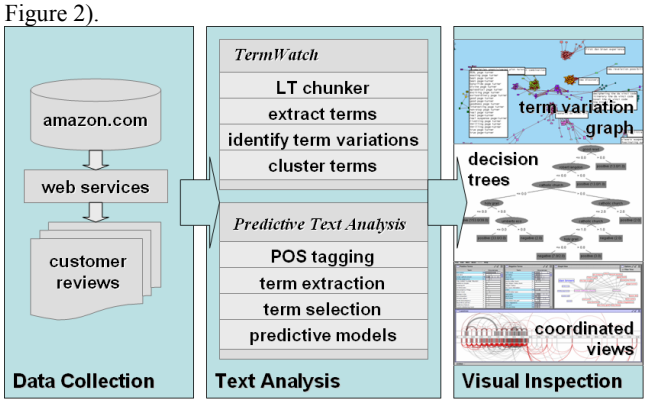

Figure 2: The overall structure of our approach.

This procedure is designed to address some of the common questions concerning conflicting opinions. *The Da Vinci Code* is a controversial bestseller. What made *The Da Vinci Code* a bestseller? Which aspects of the book are favorably reviewed? Which aspects are criticized? More generally, will we be able to apply the same technique to other bestsellers, movies, cars, electronic devices, innovations, and scientific work? Ultimately, what are the reasons and tipping points behind a success, a failure, a controversial issue, or conflicting information from multiple sources?

4.1 The Customer Review Corpus

Customer reviews of *The Da Vinci Code* were retrieved from Amazon.com using Amazon's web services (AWS). Amazon customer reviews are based on a 5-star rating system. 5 stars are the best and 1 star is the worst. In our study, reviews with 4 or 5 stars are regarded as positive reviews. Reviews with 1 or 2 stars are deemed as negative. Reviews with 3 stars are not used in this study.

The average length of positive reviews is approximately 150 words and 9 sentences, whereas negative reviews are slightly longer, 200 words and 11 sentences on average. These reviews are generally comparable to news and abstracts of scientific papers in terms of their length (See [13] for an example of a corpus of scientific abstracts).

Table 2: Statistics of the Corpus.

Corpus	Reviews	# Chars (mean)	#Words (mean)	#Sentences (mean)
Positive	2,092	1,500,707 (717.36)	322,616 (154.21)	19,740 (9.44)
Negative	1,076	1,042,696 (969.05)	221,910 (206.24)	12,767 (11.87)
Total	3,168	2,543,403	544,526	32,507

4.2 Decision Trees

The reasons we introduce the use of decision trees in our approach are twofold. First, we would like to verify the ability to predict the categories of reviews with a small set of selected terms. Second, we expect decision trees to function as an intuitive visual representation for analysts to explore and understand the role of selected terms in cateogorizing conflicting opinions. The construction of a decision tree is particular suitable for our purposes. If we use selected terms to grow a decision tree and use the review categories of positive and negative as leaf nodes, the most influential terms would appear towards the root of the tree. One would be able to explore various alternative ways to reach positive or negative reviews.

In order to put the predictive power of our decision trees in context, we generate additional predictive models of the same data with other widely used classifiers, namely the naïve Bayesian classifier and support vector machine (SVM) classifier. We expect that although decision trees may not give us the highest prediction accuracy, it should be a worthwhile trade-off given the interpretability gain.

We use the following procedure. Reviews are first processed by part-of-speech tagging. Noun phrases are extracted with stopwords removed and the last word of each term stemmed. We include adjective as part of the phrases to capture emotional and sentimental expressions. Log likelihood tests are then used to select terms that are not purely high frequent, but influential in differentiating reviews from different categories. Selected terms represent an aggressive dimensionality reduction, ranging from 94.5~99.5%. Selected terms are used for decision tree learning and classification tests with other classifiers.

We also explore the use of SVM to visualize reviews of different categories. Each review is represented as a point in a high-dimensional space $\boldsymbol{S}$, which contains three independent subspaces $\boldsymbol{S_p}$, $\boldsymbol{S_q}$, and $\boldsymbol{S_c}$: $\boldsymbol{S} = \boldsymbol{S_p} \oplus \boldsymbol{S_q} \oplus \boldsymbol{S_c}$. $\boldsymbol{S_p}$ represents a review purely by positive reviews. Similarly, $\boldsymbol{S_q}$ represents a review in negative review terms only and $\boldsymbol{S_c}$ represents. In other words, a review is decomposed into three components to reflect the presence of positive review terms, negative review terms, and terms that are common in both categories. Note that if a review does not contain any of these selected terms, then it will not have a meaningful presence in this space. All such reviews are mapped to the origin of the high-dimensional space and they are excluded from subsequent analysis.

The optimal configuration of the SVM classifier is determined by a number of parameters, which are in turn determined based on a k-fold cross-validation [17]. This process is known as model selection. A simple grid search heuristic is used to find the optimal parameters in terms of the average accuracy so as to avoid the potential overfitting problem.

4.3 Improvise

Improvise [18] is a self-contained exploratory visualization software application, written in Java and freely available on the web through an open source license. In Improvise, analysts interactively build and browse visualizations consisting of multiple coordinated views of their information. Visualizations can be rapidly modified and extended to develop hypotheses and exploit discoveries during ongoing visual analysis. In particular, Improvise provides precise control over how interaction affects the display of space, time, and abstract dimensions of information in and between multi-layer maps, scatter plots, parallel coordinate plots, tables, and other views.

An interactive visualization prototype constructed in Improvise allows exploration of time series identified by TermWatch. The main feature of this visualization is a variation on the basic two-sided arc diagram [19] in which additional time series information is displayed between the positive and negative sides. Several coordinated views allow brushing of positive and negative terms and dynamic filtering on time.

5 Descriptive Analysis of Conflicting Opinions

Table 3 shows the statistics of the term extraction and variant clustering by TermWatch. We describe these results in more detail in the following sections.

Table 3: Multi-layered feature selection using TermWatch.

Review Categories	Terms	Classes	Components	Unique Features
Positive	20,078	1,017	1,983	879
Negative	14,464	906	1,995	2,018

Figure 3 helps to identify common characteristics of positive reviews of the book. For example, many reviewers found the book a page turner, with a wide variety of minor variations, such as an amazing page turner or an episodically page turner. It indicates that the popularity of the book is in part due to its gripping plots. The ability to group these terms together is a distinct advantage for reducing the complexity of the entire terminology.

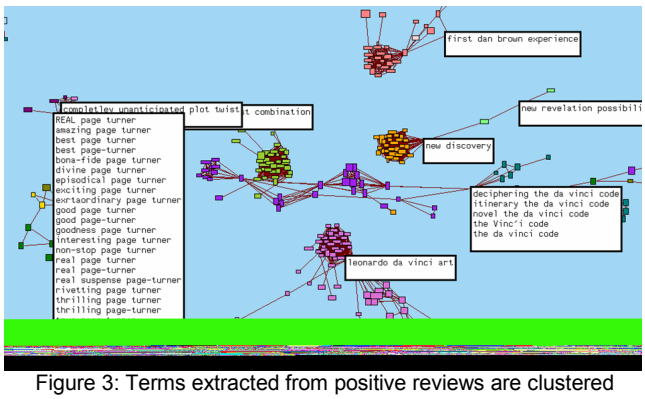

Figure 3: Terms extracted from positive reviews are clustered based on both syntactic and semantic relationships.

5.1 Analysis of reviews' thematic content

In-depth content analysis of terms in context is necessary to identify themes that differentiate positive reviews and negative reviews. The following analysis is conducted using TermWatch's navigational interface, which enables to access the complete list of terms in a cluster and the contexts in which they appeared in the corpus. A term variation network has three levels: clusters are shown at the highest level, then components, and finally terms at the lowest level.

5.2 Themes in Positive Reviews

The largest cluster labeled ***leonardo da vinci art*** in the network of terms associated with positive reviews is surrounded by the clusters *literary fiction*, *the complete dead sea scroll*, *harvard professor*, and *isaac newton*. The structure of this cluster is highly

interconnected and its content appears to be coherent as it captures the main facets of the positive reviews: comments on the major characters (*Prof Langdon*), the praises (*great storytelling, clever story, gripping novel, historic fiction*), other major characters (*Sophie Neveu, Leonardo Da Vinci, Sir Isaac Newton*). Figure 4 shows a zoom-in image of the network with the unfolded cluster shown with a pink background. There is a direct link between the component *dr robert langon* and a cluster *Havard university* which is where this professor in the book works.

Figure 4: An unfolded view of the cluster ***leonardo da vinci art*** and surrounding clusters in positive reviews.

The link towards another main cluster ***Da Vinci code fuss***, when unwrapped also shows a highly interwoven structure centered on issues about the book itself (*the da vinci code fuss, the da vinci novel, the da vinci code review*). As it turned out, such terms appeared in review titles. They were grouped into the same cluster because of the terminological variation (here modifier substitution).

The ***Da Vinci code fuss*** cluster is linked to another cluster labeled ***the vinci code***. This latter connects to another cluster labeled ***mary magdelene legend***. Unfolding the ***mary magdalene*** cluster shows that it deals with reviews arguing for the historical plausibility of events, people and organizations described in the book. For instance, there is much controversy about the supposed liaison between Mary Magdalene and Jesus Christ. Other much debated topics are the roles of the Prieure de Sion and Opus Dei organizations, the effects of the historical events as depicted in the book on religious faith of today's Christians, the research the author claimed to have carried out to back up his versions of the historical events. Because of the varied nature of the terms in this cluster, most of the links are due to associations (co-occurrence).

An isolated sub-network deals with the author's writing track: his next, previous or new books. Apparently, the terminology used to talk about this in the reviews is distinct from the terms used to praise the current book, hence the isolation. More clusters featured in the positive reviews network but for space reasons, we cannot analyze all of them here.

5.3 Themes in Negative Reviews

We took an in-depth look at the context of occurrence of the most frequent terms in negative reviews, namely *mary magdalene*, *opus dei*, *the holy grail*, *too much*, *art history*, *good book*, *page turner*, *secret society*, *the last supper*, *conspiracy theory*, and *villain*. We found out that the negative reviews questioned the historical and religious foundations of the books which the author (Dan Brown) presented as "truth based on research." The author's claims came under ferocious criticisms by the negative reviewers who undertook to prove point by point that the author is an impostor. The most controversial point is centered on the religious facts portrayed in the book such as the supposed love affair and subsequent marriage between Jesus Christ and Mary Magdalene. Indeed, the term *mary magdalene* is consistently featured in all the negative reviews from the first year since the book was published in March 2003. Table 4 shows an example of a negative review containing this term. The visualization obtained from the chronological analysis of negative reviews further illustrates this point (See Figure 6).

Table 4: Example of a negative review containing the term "mary magdalene."

Title: *Gripping, but definitely "fiction"* Date: *2004-03-17* Rating: *1* Review: *I know you've read some incredible things about* mary magdalene *and her fling with Jesus the Christ in THE DA VINCI CODE by educated Harvard writer Dan Brown. Many of these theories come out of a well-financed (Hollywood financed!) minority of revisionist scholars whom the press sees as more exciting when they are, in fact, just speculating.*

5.4 Time series analysis of terminological evolution

The timing of the creation of a term variation link is of particular interest to us because we want to identify when a significant terminology change takes place. Figures 5 and 6 show timestamps on term variation links. A timestamp is linked to a term through a yellow dashed line. By switching back and forth between term variation links and time stamped links one can narrow down the time frame in which terms are associated with. In Figure 5, the more yellow lines a term is associated with, the more persistent the term in the positive reviews. In contrast, the appearances of terms with few yellow lines are sporadic in the corpus. Therefore, persistently occurring terms are placed in the core of the network and they are connected through many yellow lines.

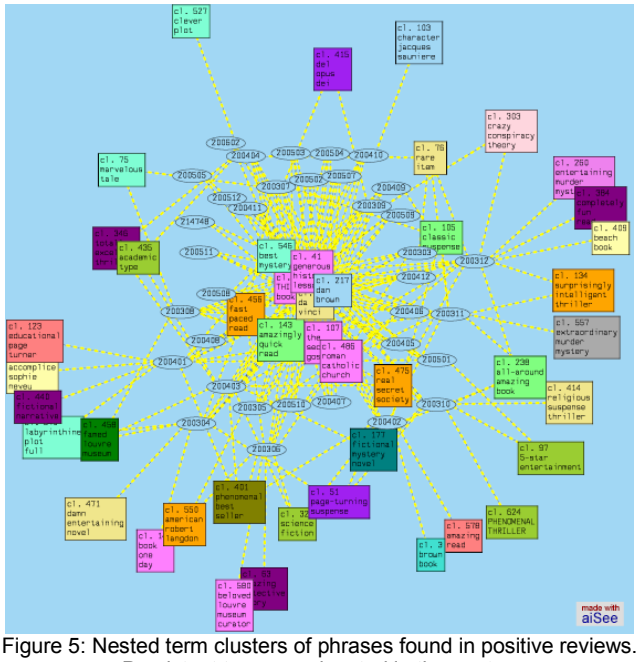

Figure 5: Nested term clusters of phrases found in positive reviews. Persistent terms are located in the center.

The most persistent themes in this graph are *dan brown, da vinci, generous history lesson, best mystery, amazingly quick read, roman catholic church, page turning suspense,* and *fast paced read*. These themes echo the sentiments of people who loved the book. The two main reasons are historical and fictional. Some found it well written and the other loved the historical background. It is worth noting that the detractors of the book also used the same two main reasons (style of writing and historical background) to demolish the book. This can be verified against

the histogram generated for terms used in positive and negative reviews (See Figure 7).

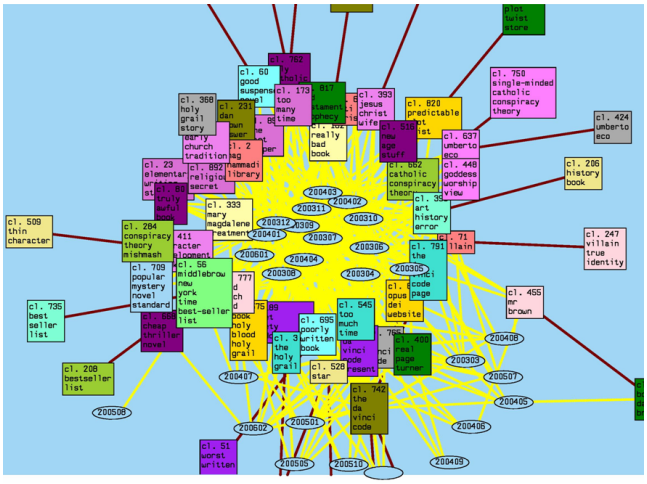

Figure 6: Nested term clusters of phrases found in negative reviews.

In contrast, the graph generated for the chronological analysis of term variants in negative reviews did not have a core; terms formed a circular shape instead (See Figure 6). There are many more links, however. Some of the themes portrayed by the clusters clearly have negative connotations, for example, *really bad book, poorly written book, truly awful book, catholic theory conspiracy, art history error, conspiracy theory mismatch,* and *thin character*.

An additional visualization tool is developed so that users can explore terms found in positive reviews alongside terms found in negative reviews in the same month. Positive reviews appear to have fewer terms than negative reviews, although there are more positive reviews than negative ones.

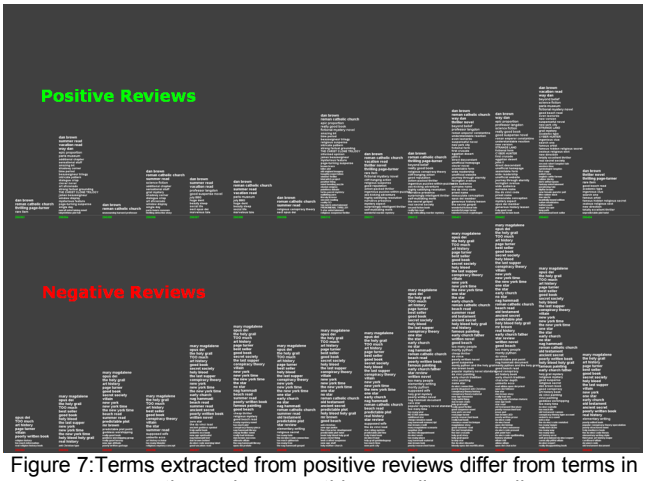

Figure 7:Terms extracted from positive reviews differ from terms in negative reviews monthly as well as overall.

Figure 7 shows a steady growth in the number of terms from negative reviews over the 13 months. However, such a steady growth is absent from the positive review timeline. It is our observation that positive reviews may need a few well-chosen adjectives to express their enthusiasm as well as commenting generally on the plot while negative reviewers have to do extensive research in order to challenge the book point by point. Thus, negative reviews, on the average, tend to be longer than positive ones. A similar observation can be made to the reviews of scientific papers, where negative reviews tend to be more detailed than positive ones, but we are not aware of empirical data to verify this observation.

5.5 Coordinated Views of Terms

Figure 8 shows a screenshot of the coordinated views created by Improvise. In the modified arc diagram, shown in the lower half of the screen, increasing time is labeled from left to right along the center axis. The top half of the diagram shows terms used in positive review, the bottom half terms used in negative ones. Arcs connect months in which common terms appear. Arc thickness represents the number of common terms. Bar thickness shows the number of terms for each month, considered individually. (Bar thickness tends to be thicker than arc thickness, because some terms appear in only one month.)

Two multicolumn tables provide detail about positive and negative terms. For each term, a nested slider (navigationally coordinated with the arc diagram) shows the pattern of monthly occurrences as a simple time series. A graph shows selected terms as nodes (blue for positive, red for negative, magenta for mixed). Node size encodes total appearances of each term. Edges connect terms that appear the same month, with thickness representing the number of months in common. Selected terms are highlighted in red in the arc diagram. If the time filter checkbox is selected, the tables and graph filter out terms for months outside the time range visible in the arc diagram. Analysts can brush interesting terms in any of the views, explore patterns of term usage by panning and zooming over time, then drill down to compare temporal patterns for particular terms.

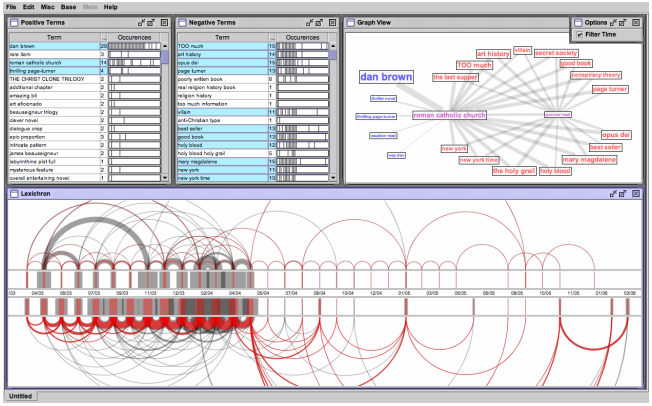

Figure 8: Coordinated views generated by Improvise.

6 Predictive Text Analysis

The predictive text analysis of the book reviews serves two objectives: to validate the predictive power of selected terms and to provide a visual representation for analysts to explore and understand the role of these terms in reviews of different categories.

Table 5 summarizes the number of terms selected by log likelihood values and the accuracies of three classifiers with 10-fold cross-validation. The original set of extracted terms contains 28,763 terms. The dimensionality reduction rates range from 94% to 99.5%. In contrast, if we select terms based on their document frequencies (>=2), there will be 6,881 terms and the accuracy of classification if a C4.5 decision tree is 68.89%, which is below all the models with log likelihood tests. More importantly, decision trees (C4.5) are relatively stable in terms of 10-fold cross-validation accuracies (slightly over 70%), whereas SVM models are more than 80% of accuracy, which means the selected terms are good candidates to categorize these reviews. Classifiers used in this study are available in Weka [20].

Table 5: Classification accuracy on 10-fold cross-validation.

Log Likelihood (p-level)	Selected Terms	C4.5	NaiveBayes	SVM
0.05	1,666	70.26	77.54	84.59
0.01	360	71.67	76.67	83.14
0.001	146	70.01	75.74	81.72
Doc Freq (>=2)	6,881	68.89		

Although using document frequencies as feature selection metrics may give comparable results to metrics such as information gain, it is not as efficient as other metrics if aggressive dimensionality reduction is desired [21]. Terms selected by log likelihood tests of the presence and absence of a term in relation to the category of a review are visualized in Figure 9 with GGobi[4]. It shows the majority of terms have relatively low document frequencies. On the other hand, terms such as *money*, *hype*, *character*, and *great read* are selected with quite different document frequencies.

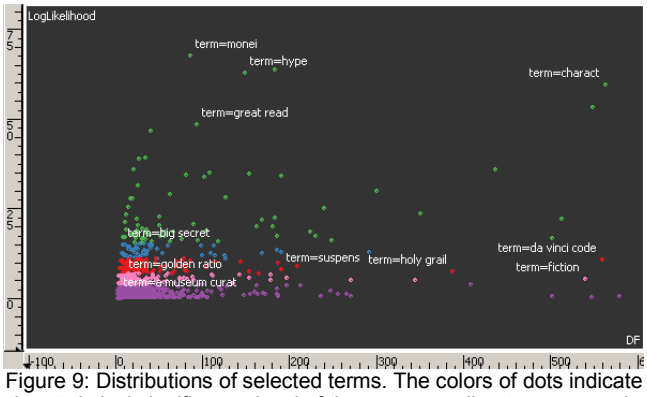

Figure 9: Distributions of selected terms. The colors of dots indicate the statistical significance level of the corresponding terms, namely green (< 0.001), blue (p=0.001), red (=0.01), and pink(=0.5).

6.1 Decision Trees

Two decision trees are shown in Figures 10 and 11 to illustrate how they may facilitate an understanding of conflicting opinions. The top of the tree contains terms that strongly predict the category of a review, whereas terms located in lower part of the tree are relatively weaker predictors.

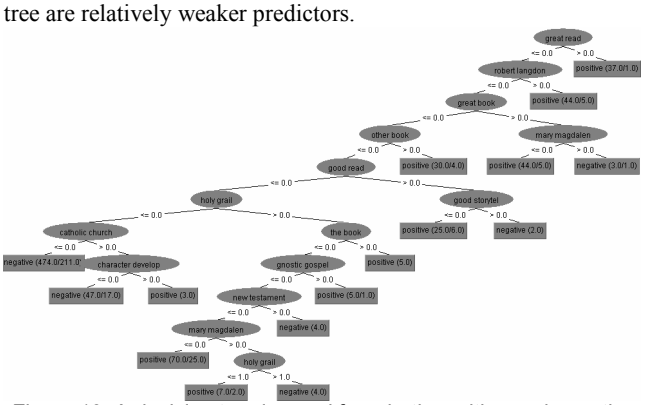

Figure 10: A decision tree learned from both positive and negative reviews in 2003. Terms are stemmed.

In the 2003 decision tree, for example, the presence of term *first page* predicts a positive review. The term *good read* is also quite obvious. Interestingly, the route *robert langdon* ➔ *jesus christ* predicts a negative review. The analyst could then form a hypothesis that some negative reviews are probably to do with these two terms. Similarly, the branch at the lower right corner shows that *mary magdalen*➔ *holy grail* also leads to negative reviews. The 2004 decision tree is dominated by the term gread read. Terms that are closely connected to negative reviews include *mary magdalen*, *new testament*, and *holy grail*.

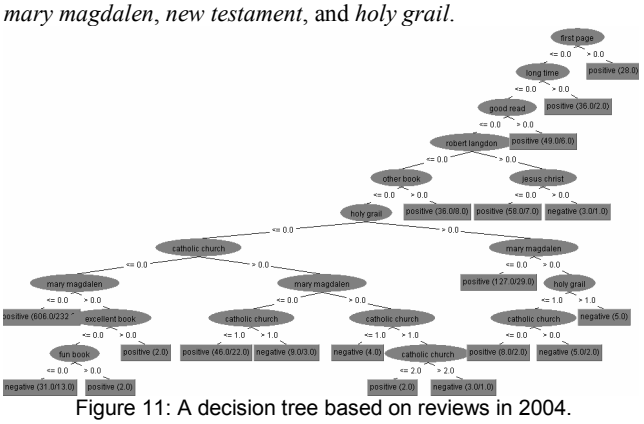

Figure 11: A decision tree based on reviews in 2004.

6.2 Classifying Reviews by Active Terms

Both positive and negative reviews in this dataset contain a large number of terms. Linguistically active terms in TermWatch are terms that have many variants. Active terms are used to label the clusters and they represent a much smaller portion of the phrases, which is 8.3% of the noun phrases extracted by the LT-chunker, a component of LTPOS.

Since these active terms have not been selected based on their occurrence in the reviews, they are not expected to be the best candidates for indexing reviews in a classification task. We index reviews with TermWatch cluster labels if terms in a cluster appear in reviews. These terms are expected to be closely related to cluster labels as they are linked by a short chain of variations. We then generated an additional decision tree using 60% of the data as a training set. The resultant decision tree correctly classifies 68% of the remaining reviews. This accuracy is lower than those obtained by previous classifiers, but it remains interesting since it is mainly based on long multiword terms, which tend to have much lower frequencies than single-word terms.

The accuracy of this decision tree relies on the 30 most active terms, i.e. they have the greatest number of variants in reviews. For example, *robert langdon story* has 250 variants in reviews of which 85% are positive. Similarly, terms such as *opus dei website*, *millennium-old secret society* and *historical fact revelation* have more than 100 variants in reviews of which 66% are positive.

Browsing terms not included in the decision tree model is also informative. For example, each of the terms *anti christian*, *secret grail society blah blah blah*, and *catholic conspiracy* has only 6 variants in reviews, all negative, identifying readers shocked by the book.

Browsing the interrelationship between reviews and TermWatch clusters reveals topics that appear in both categories positive/negative and thus ignored by decision trees. As it turns out, each of the terms like *jesus christ wife*, *mary magdalene gospel*, *conspiracy theory* and *christian history* have more than 50 variants that are almost evenly distributed between positive and negative reviews.

7 Discussions Discussions and Conclusions

The microscopic level visual analysis has identified some salient features that discriminate between positive and negative reviews. Such features play a fundamental role in sense making involving diverse perspectives, conflicting opinions, or contradicting evidence. The term variation focus has made it

[4] http://www.ggobi.org/

relatively straightforward to identify the predominating themes of positive and negative reviews. For negative reviews, the heavy religious controversies raised by the book are signified by a set of persistent and variation rich terms such as *mary madgalena, opus dei,* and *the holy grail*, and none of these terms ever reached the same status in positive reviews. Much of the enthusiasm in positive reviews can be explained by the perspective that the book is a work of fiction rather than scholarly work with discriminating terms such as *vacation read, beach read,* and *summer read*.

To our knowledge this is the first visual analytics example of conflicting book reviews over an extensive period of time. We are encouraged by the initial results. This study has also identified challenges and research questions that need to be pursued further. For example, what insights would we gain if we were using traditional statistical-oriented and high-frequency-focused approaches? Are there potential biases introduced by exclusively focusing on term variation patterns? How does the term-level microscopic perspective complement with topic-level or domain-level macroscopic visualizations of the dynamics of thematic evolution?

In conclusion, we found the combination of term variation patterns, statistical tests of associations, predictive models, and various interactive visualization tools a promising and generic approach to visual analysis of conflicting views, especially in feature selection and handling low-frequency but critical connections. The use of various coordinated visualizations is necessary when multiple levels of abstraction and multiple perspectives are involved. Comprehensive evaluations of alternative approaches and thorough investigations of the applicability in a wider range of visual analytic tasks are among the most important issues to be addressed in further studies.

Acknowledgements
The work is in part supported by the Northeast Visualization and Analytics Center (NEVAC).
http://www.geovista.psu.edu/NEVAC/index.html

REFERENCES

[1] J. J. Thomas and K. A. Cook, "Illuminating the Path: The Research and Development Agenda for Visual Analytics," IEEE Computer Society Press, 2005.

[2] M. Q. W. Baldonado and T. Winograd, "SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests," Proceedings of the SIGCHI conference on Human factors in computing systems, Atlanta, Georgia, 1997, pp. 11-18.

[3] C. Plaisant, J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord, "Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces," Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL'06), Chapel Hill, North Carolina, 2006, pp. 141-150.

[4] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims, "Identifying Temporal Patterns and Key Players in Document Collections," Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05), 2005, pp. 165 – 174.

[5] T. K. Landauer, D. Laham, and M. Derr, "From paragraph to graph: Latent semantic analysis for information visualization," *PNAS*, vol. 101, pp. 5214-5219, 2004.

[6] N. S. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: Automated Trend Discovery for Weblogs," WWW2004, New York, NY, 2004.

[7] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 9-20, 2002.

[8] A. Kaban and M. A. Girolami, "A dynamic probabilistic model to visualise topic evolution in text streams," *Journal of Intelligent Information Systems*, vol. 18, pp. 107-125, 2002.

[9] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proceedings of the ACL, 2004.

[10] C. Jacquemin and D. Bourigault, "Term extraction and automatic indexing," in *The Oxford Handbook of Computational Linguistics*, R. Mitkov, Ed. Oxford, England: Oxford University Press, 2003, pp. 599-615.

[11] B. Daille, "Conceptual structuring through term variations," Proceedings of the ACL-2003 Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment, Saporro, Japan, 2003, pp. 9-16.

[12] C. Fellbaum, *WordNet: An electronic lexical database*. Cambridge, MA.: MIT Press, 1998.

[13] F. Ibekwe-SanJuan and E. SanJuan, "Mining textual data through term variant clustering: The TermWatch system," Recherche d'Information assistée par ordinateur(RIAO 2004), University of Avignon, France, 2004, pp. 487-503.

[14] F. Ibekwe-SanJuan, "A linguistic and mathematical method for mapping thematic trends from texts," Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98). , Brighton, UK, 1998, pp. 170-174.

[15] L. Lebert, A. Salem, and L. Berry, *Exploring textual data*. Boston, Massachusetts: Kluwer, 1998.

[16] E. SanJuan and F. Ibekwe-SanJuan, "Text mining without document context," *Information Processing & Management*, 2006.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[18] K. Weaver, "Magnesium and Migraine," *Headache*, vol. 30, pp. 168-168, 1990.

[19] M. Wattenberg, "Arc diagrams: Visualizing structure in strings " Proceedings of the IEEE Symposium on Information Visualization, Boston, MA, 2002, pp. 110 – 116.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufmann, 1999.

[21] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," The 14th International Conference on Machine Learning (ICML'97), Nashville, US, 1997, pp. 412-420.