Visual Analysis of Conflicting Opinions.

Chaomei Chen, Fidelia Ibekwe-Sanjuan, Eric Sanjuan, Chris Weaver

▶ **To cite this version:**

Chaomei Chen, Fidelia Ibekwe-Sanjuan, Eric Sanjuan, Chris Weaver. Visual Analysis of Conflicting Opinions.. IEEE Symposium on Visual Analytics Science and Technology 2006, Oct 2006, Baltimore, Maryland, United States. pp.319-330. hal-00636138

HAL Id: hal-00636138
https://hal.science/hal-00636138

Submitted on 26 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Visual Analysis of Conflicting Opinions

[1]Chaomei Chen, Drexel University, USA

[2] Fidelia Ibekwe-SanJuan, Université Jean Moulin, France

[3]Eric SanJuan, Université de Metz, France

[4]Chris Weaver, Penn State University, USA

ABSTRACT

Understanding the nature and dynamics of conflicting opinions in texts is a challenging issue concerning sense making and decision making in visual analytics. In this paper we address this issue with a visual analysis of how positive and negative reviews of the controversial bestseller *The Da Vinci Code* differ. An integrative approach is proposed to capture the dynamics of opinions at both macroscopic and microscopic levels. More than 3,000 customer reviews of the book retrieved from Amazon.com, including 1,738 positive and 918 negative reviews, are analyzed in this approach to address questions such as: what are the differentiating features of positive and negative reviews? How did positive and negative reviews evolve through time? To what extent can discriminating features be automatically extracted without prior knowledge of the subject? To what extent can these features accurately predict the general opinion of a given review? In order to identify differentiating features, we focus on underlying terminology variations and use TermWatch, a tool for terminology variation analysis, to construct multi-layered networks of terms based on syntactic, semantic, and statistic associations. Time series of terms with strong variations are constructed on a monthly basis. We have utilized a number of visualization and modeling tools to conduct a visual analysis of these identified features. The results show not only what positive and negative reviews have in common, but also what differentiate them persistently over time. Furthermore, how earlier reviews may have influenced subsequent reviews are also investigated. Challenges for future work are identified.

CR Categories and Subject Descriptors: Visual Analytics, Information analytics, Text and Document Visualization

Additional Keywords: analysis of conflicting information, visualization of terminology change, terminology variation, book reviews.

1 INTRODUCTION

Understanding the nature and dynamics of conflicting opinions in texts is a challenging issue concerning sense making and decision making in visual analytics. Contradictory opinions exist in a diverse range of scientific, engineering, social, political, biomedical domains. Key challenging tasks include identifying the basic premises of arguments, assessing the credibility of existing evidence, understanding the context and background of a particular position, and tracking the dynamics of how various opinions evolve and how they interact with a broader context of information. The ability to accomplish such tasks effectively and efficiently has a direct impact on people's understanding and decision making processes. 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].

A bestseller book such as *The Da Vinci Code* can reach tens of millions of readers. Reviews of the book by a diverse group of readers provide a valuable source of insight in terms of how people form their opinions and what might influence their opinions. In this paper, we present a study of such reviews as a vehicle to improve our understanding of underlying technical challenges for understanding contradicting opinions expressed in a number of genres. Choosing this topic has some distinct advantages: no prior domain knowledge required, easy to interpret and evaluate results, potentially extensible applications to other genres.

Figure 1 depicts the number 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 outnumbered negative ones, arguments and reasons behind these reviews are not apparent.

Figure 1. 1,738 positive (green) and 918 negative (red) reviews of *The Da Vinci Code*. Positive reviews have customer ratings of 4 or 5. Negative reviews have ratings of 1 or 2.

Indeed, the questions concerning book reviews have implications beyond books. One may ask similar questions about merchandise, electronic devices, information services, or opinions on wars, religious, and environmental issues. One also needs to address issues concerning the underlying credibility of evidence, the strength of arguments, diverse perspectives, and expectations.

2 RELATED WORK

A few areas of research are relevant to the research issue we intend to address, namely discovering trends in streams of information from a diverse range of sources, identifying the diffusion of ideas and tipping points, and analyzing sentiments in movie reviews.

1. chaomei.chen@cis.drexel.edu
2. ibekwe@univ-lyon3.fr
3. eric.sanjuan@iut.univ-metz.fr
4. cew15@psu.edu

Kleinberg [2] identified three predominating approaches to the study of temporal dynamics of information streams, namely threshold-based, state-based, and trend-based methods. These techniques have been applied to the analysis of weblogs, Internet search queries, and usage data at high-traffic websites. He identified the problem of alignment between patterns identified in a data stream and underlying events in the real world.

Identifying the impact of scientific publications has been the subject of a variety of research. For example, citation analysis aims to identify the impact of a publication in terms of the number of citations it receives and how paradigm shifts manifest themselves in scientific literature [3]. In contrast to the use of references as indicators of intellectual impact, researchers have also attempted to identify influential papers without the presence of references [4]. For instance, a simple approach is to group text documents into clusters of similar topics. Since documents within each cluster can be arranged chronologically, it is reasonable to assume that papers positioned in the earlier years contribute to the thematic development of the given cluster. However, chronologically leading is one thing, semantically influencing may be another. Research in this area still faces fundamental challenges.

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 [5], 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.

Sentiment analysis is another area of relevant research, which aims to identify underlying viewpoints based on sentimental expressions in texts. Pang and Lee [6] 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.

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.

The majority of relevant research is built on the assumption that desirable patterns are statistically detectable. 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, our approach is built on research in terminology variation and combines with interactive visualization functions to support the understanding, interpretation, and verification of conflicting information from a diverse range of perspectives. We demonstrate our approach with an example of analysis of customer reviews of the controversial bestseller *The Da Vinci Code*. In particular, 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 [9]. It focuses on symbolic relations between terms and how they can be related through several types of variations and transformations. Research in the computational terminology community has established the importance of variation phenomena amongst domain terms [10].

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, for example, WordNet [11]. 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.

Table 1. Linguistic operations for 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 [12, 13] is originally designed to monitor the development of scientific and technological domains. It combines surface linguistic analysis with a scalable clustering algorithm in order to visualize the important topics contained in a text corpus. This lexico-syntactic approach is suitable for clustering multi-word terms (MWTs) which rarely re-occur in the texts. MWTs 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[1]. Then, the different terminological variations are identified between terms. These identified terms variants are

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

subsequently clustered with a hierarchical clustering algorithm CPCL (Classification by Preferential Clustered Link).

The CPCL algorithm operates in two stages. First, connected components are formed based on semantically related terms through a subset of the term variation types called COMP. These components typically include spelling variants, WordNet semantic variants, and modifier variations. The idea is to group together conceptually related terms, which only differ by their modifiers but share a common head word.

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. 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 [14]. The clustering algorithm is implemented using a straightforward $\boldsymbol{O}(E)$ procedure called SLME (*Select Local Maximum Edge*) described in [15].

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

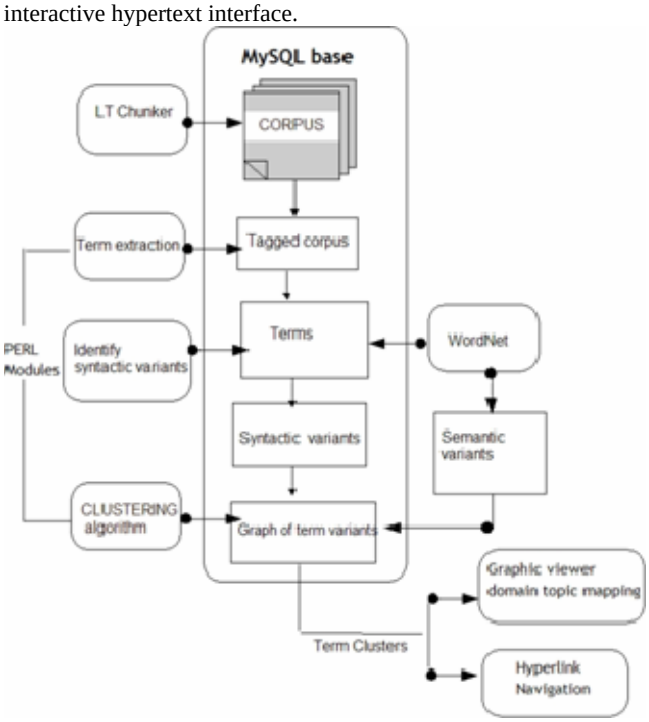

Figure 2. Overview of the *TermWatch* system.

The clusters of term variants generated by TermWatch have properties that 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 clustering is done on connected components but not on terms, it is necessary to be able to unfold a cluster down to the term level. The *aiSee*[2] 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 show its internal structure: the connected components, the most active variants. The user can explore and study 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.

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? What are the reasons given by positive customer reviews on Amazon.com? What are the reasons in negative reviews? More generally, will we be able to apply the same technique to other bestsellers, movies, cars, electronic devices, innovations, and scientific work? In general, what are the reasons 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 service. Amazon customer reviews are based on a 5-star rating system. 5 stars are the best rating and 1 star is the worst. Reviews with 4 or 5 stars are regarded as positive reviews in our study. Reviews with 1 or 2 stars are deemed as negative. Reviews with 3 stars are not used in the analysis.

Positive reviews are approximately 150 words long and 9 sentences on average, 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 [12] 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 Improvise

Improvise [16] 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

2 http://www.aisee.com.

in and between multi-layer maps, scatterplots, 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 [17] 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. (See Figure 5 for more details).

4.3 Support Vector Machine Classifier

In order to evaluate the extent to which these terms selected based on term variation types capture the overall judgmental opinion of a review, we developed a predictive model of customer reviews based on these terms using support vector machine (SVM). SVM is a widely used and powerful machine learning technique for classification [18]. Each book review is labeled as positive or negative. The first 250 positive reviews and the first 250 negative reviews are used as the training set to build an SVM classifier. The classifier is then applied to the rest of the reviews to evaluate the accuracy of the model.

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 [19]. 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.

5 Visual Analysis of Conflicting Opinions

Selecting variation relations is effectively 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.

Table 3. Multi-layered feature selection using TermWatch.

Review Categories	Terms	Classes	Components	Unique Features
Positive	2,0078	1,017	1,983	879
Negative	1,4464	906	1,995	2,018

5.1 Term Variation Networks

Figure 3 is generated 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 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. The following figure shows the 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 timeframe in which terms are associated with. In Figure 4, 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. Details of these terms can be checked against the histogram generated for terms used in positive and negative reviews (See Figures 6 and 7).

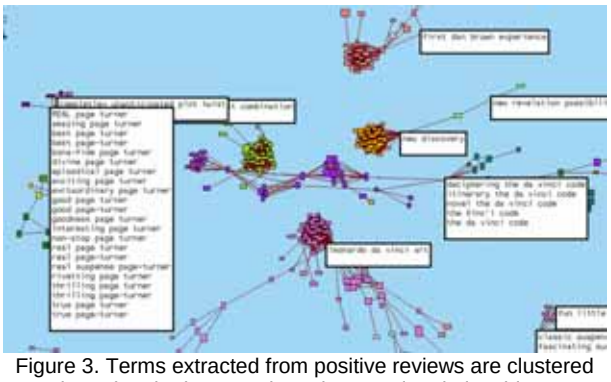

Figure 3. Terms extracted from positive reviews are clustered based on both syntactic and semantic relationships.

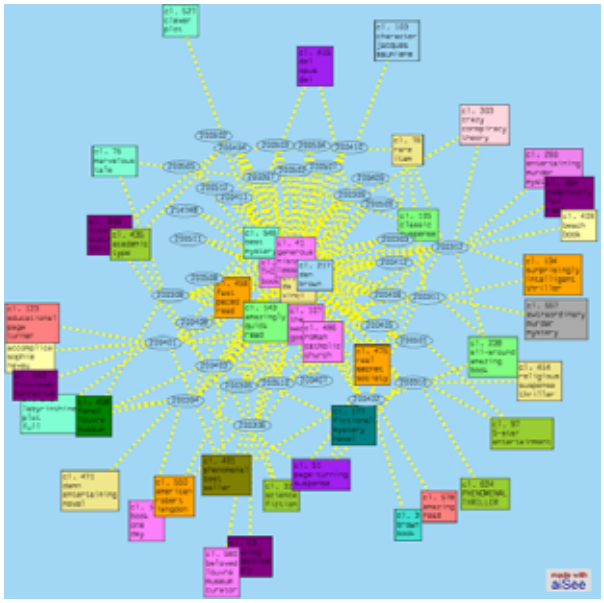

Figure 4. Nested term clusters of phrases found in positive reviews. Dashed yellow lines link clusters to timestamps of their occurrences.

5.2 Coordinated Views of Terms

Figure 5 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.

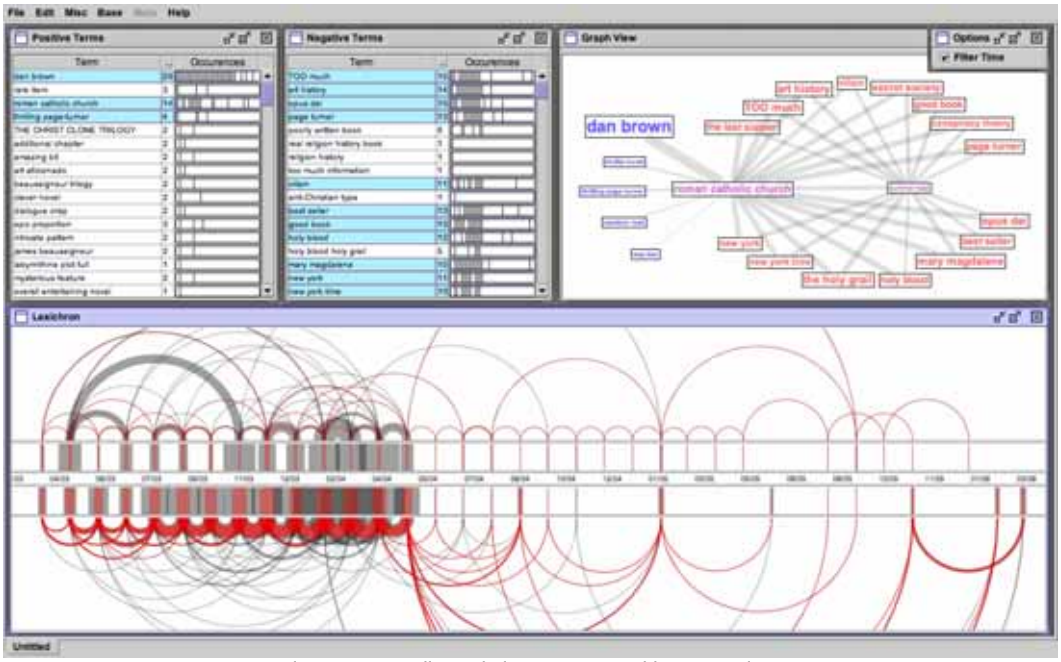

Figure 5. Coordinated views generated by Improvise.

5.3 Time Series of Differentiating Terms

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.

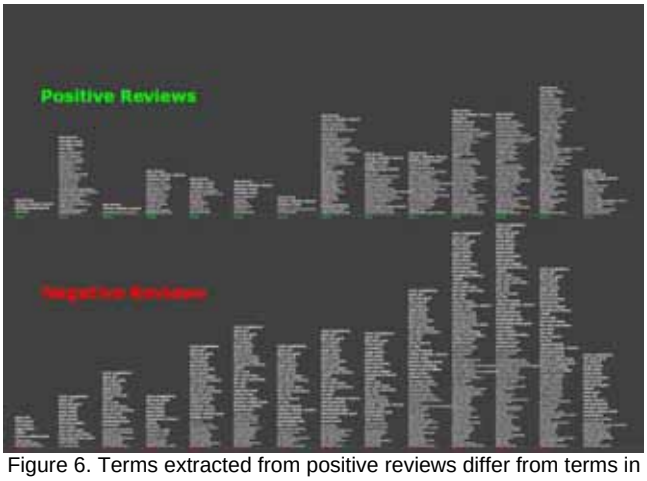

Figure 6. Terms extracted from positive reviews differ from terms in negative reviews monthly as well as overall. See also Figure 7.

There is 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.

6 Content Analysis of Persistent Themes

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.

6.1 Themes in Positive Reviews

The largest cluster "**Leonardo Da Vinci art**" in the network of terms associated with positive reviews is surrounded by the clusters "*Literary fiction*" (a qualifier of the book), "the complete dead sea scroll" (a book on dead/lost secret society), the cluster "*Harvard professor*" (the main character of the book), "*Isaac Newton*" (an important part of the plot). If we unfold this cluster, we can see the structure of the components is highly interconnected.

The structure of the core 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*), this last one being linked to the core cluster.

The link towards another main cluster "**Da Vinci code fuss**", when unwrapped also shows a very 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 and the terminological variation (here modifier substitution) enabled to group them in one cluster.

This cluster is linked to another called "**the vinc'i code**'' which has a part of components on this name and another part, based on co-occurrence relations with components like *rubber ducky*, *New York city*, *battery park*, *English cattle farmer*. Trying to understand what this new vocabulary meant, we traced them back to the reviews where the terms appeared. *Rubber ducky* figured in two reviews give a different account of the book from the mainstream positive reviews. This cluster is linked on one hand to a small cluster named "**Da Vinci code beauty**" and on the other to a large cluster called "**Mary Magdelena legend**".

Unfolding the "***Mary Magdalena cluster***" reveals that it deals with reviews arguing the historical plausibility of events, people and organizations evoked in the book. For instance, there is much controversy about the supposed liaison between Mary Magdalena and Jesus. 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).

Upon unfolding the "**the Vinci's code**" cluster, there appears to be 2 sub-clusters: one formed by variation relations around this term (lexical inclusions). They are mostly longer terms containing this generic term) and another formed by association relations (co-occurrence).

A detached 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 of this sub-network. Upon checking, most of the links between terms in these components are linguistic (variations), which explains why the network is isolated.

6.2 Themes in Negative Reviews

We took an in-depth look at the context of occurrence of some negative terms (mary magdalene, opus dei, the holy grail, too much, art history, good book, page turner, secret society, the last supper, conspiracy theory, villain).

Looking at the context of appearance of the topmost negative terms, one finds that the negative reviews were focused on the historical and religious foundations of the books which the author (Dan Brown) presented as "truth based on research". The author ' s claims come under ferocious criticisms by the negative reviewers who undertake to prove point by point that the author is an imposter. The most controversial point is centered around the religious facts portrayed in the book such as the supposed love affair and subsequent marriage between Jesus Christ and Mary Magdalene. Indeed, this term figured consistently in all the negative reviews from the first year since the book was published in March 2003. The following table shows a few negative reviews containing this term.

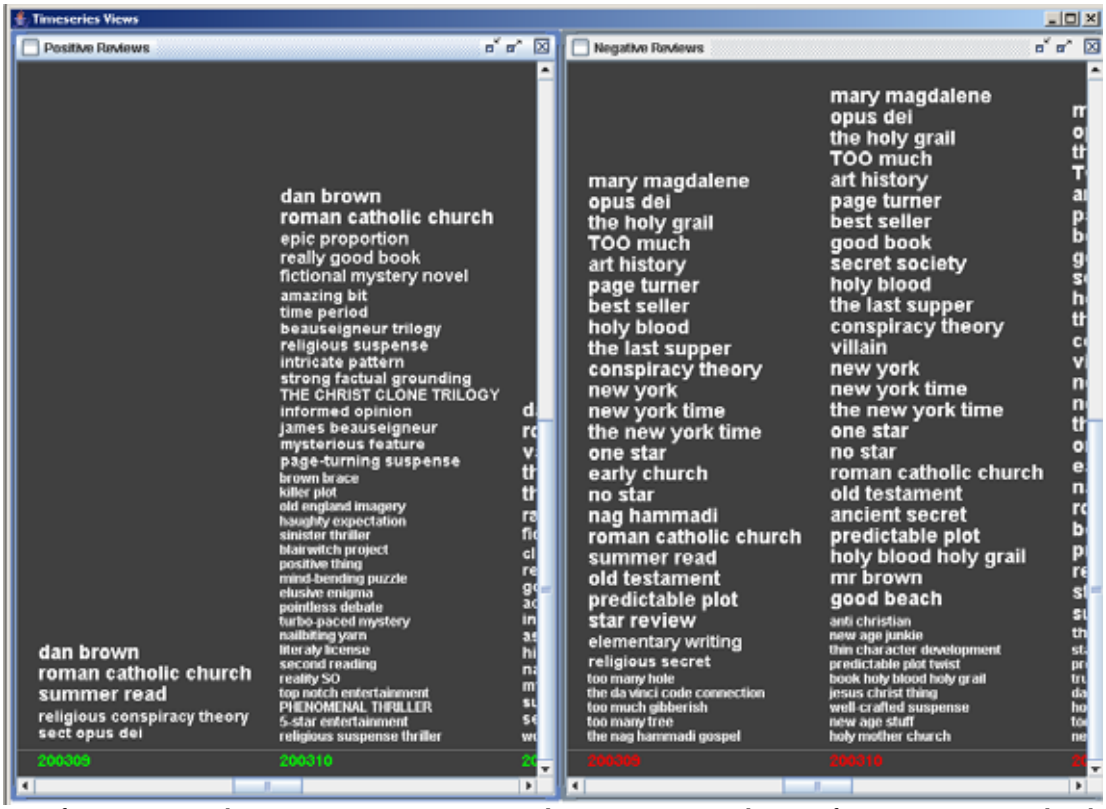

Figure 7. List of active terms with strong variations in positive and negative reviews. The size of a term is proportional to the number of months the term appeared in reviews. The most popular terms are placed on the top of each month's list. The least popular ones are at the bottom of the list.

Table 4. Examples of negative reviews containing the term "mary magdalene."

Bad Fiction, Bad History 2004-04-08 Rating: 1 Our Junior Batmen are chasing after the Holy Grail, which in Brown's universe, shaped as it is by popular conspiracy-theory speculations rather than certified scholarship, is not the cup of Christ, but a "royal bloodline" composed of descendants of Jesus Christ and (who else?) **mary magdalene**. This theory has been promoted without success before, most notably in the 1983 book Holy Blood, Holy Grail by Michael Baigent, Henry Lincoln, and Richard Leigh (New York: Dell). That book has been soundly critiqued.
Gripping, but definitely "fiction" 2004-03-17 Rating: 1 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.
illogical, inconsistent and inaccurate 2004-03-08 Rating: 1 I agree that throughout the history of the Church, the feminine aspect of the Divine has been suppressed. But Brown's suggestion that this is embodied in **mary magdalene** (if such a person in fact ever existed), perpetuated by a cult that has to be kept in secrecy for ages is not logical. There are other points: such as his theory about the bloodline of Jesus (why France?), Mary instead of John at the Last Supper, **mary magdalene** as the Holy Grail, even the symbolism in Disney's Lion King (a real long stretch) -- all thrown together to make a good story.
A Borrowed Idea, So-So Writing, and a Gaping Plot Hole 2004-02-18 Rating: 2 As other reviews have pointed out, the basic foundation for this novel is taken straight from the idea that the Holy Grail is actually **mary magdalene** in her role as Jesus's wife, carrying his child.

7 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. These terms 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.

Figure 8 shows the trace of the grid search we used for model selection. Two major parameters c and g are chosen at the average 65.47% of accuracy. The selected parameters are then used to train the SVM classifier and test it on remaining reviews.

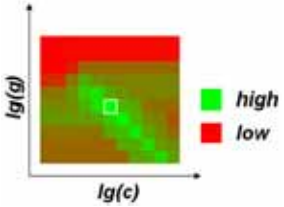

Figure 8. Model selection based on grid search. The highest cross-validation value is 65.47.

The SVM model classified reviews into positive and negative categories with 67.14% of accuracy. Considering that the construction of the model is only based on 8.3% of the entire set of terms, essentially noun phrases, this level of accuracy suggests that terms selected by term variation relations play an important role in differentiating conflicting opinions.

The performance of the SVM model can be partially visualized by projecting reviews from the high-dimensional space to a unit square in a 2-dimensional space. The position of a review in the 2-dimensional space (x, y) corresponds to the ratio of the number of positive or negative terms found in the review to the total number of base terms found in it.

Figure 9 shows the visualization. Each dot in the image represents a review. A blue dot represents a positive review, whereas a yellow dot represents a negative review. Blue dots in the lower right quadrant would be positive reviews that used many terms identified by TermWatch. Yellow dots in the upper left quadrant would be negative reviews that used many negative terms. The vertical line of yellow dots along the left edge of the image corresponds to a group of negative reviews that used a lot negative terms but none of the positive terms. The blue area is the predicted area of positive reviews based on their simplified 2-dimensional representations. It is clear further studies are needed to better understand the role of term variation in predicting the category of a review, but the results appear to be encouraging because we are able to identify terms that characterize the underlying differences between conflicting reviews without assuming any prior knowledge of the subject. Therefore we expect this approach has the potential to be applicable to a wider range of information involving multiple sources and multiple perspectives.

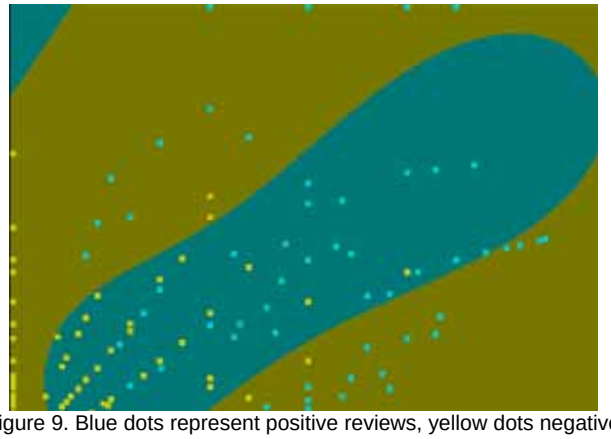

Figure 9. Blue dots represent positive reviews, yellow dots negative reviews. The blue area identifies the boundaries predicted by the SVM model for positive reviews.

8 Discussions and Conclusions

TermWatch found 20,078 terms from a corpus of 1,733 short texts, which may appear to be too much at a first glance. Fortunately, there is a linguistic term filtering mechanism in TermWatch, via the variation relations. Only those terms that share some variation relations (and now also enough co-occurrence relations) appear in components/classes.

This is the first time TermWatch is applied to a general English corpus. Until now, it has only been run on scientific and technical English. The linguistic patterns defined for term extraction and the variation relations withstand the confrontation with everyday English. The corpus of book reviews is contributed by a diverse population of readers on general topics. TermWatch was designed to extract domain terms, meaning some specialized vocabulary. In this study, we intentionally have not made any assumptions about the underlying vocabulary, and the approach still identified meaningful variations and form coherent clusters.

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 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, 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, 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-biased 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 term variation a good, generic candidate for visual analysis of conflicting views, especially in feature selection and handling low-frequency but critical connections. The use of various visualizations is necessary when multiple levels of abstraction and multiple perspectives are involved. The use of SVM has the potential to provide an informative evaluation framework. Comparisons to alternative approaches and thorough investigations of the applicability in a wider range of visual analytic tasks will make important contributions to the visual analytics research.

Acknowledgements

Chaomei Chen wishes to acknowledge the support of Northeast Visualization and Analytics Center (NEVAC).

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] J. Kleinberg, "Temporal dynamics of on-line information streams," in *Data Stream Management: Processing High-Speed Data Streams*, M. Garofalakis, J. Gehrke, and R. Rastogi, Eds.: Springer, 2005.

[3] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 359-377, 2006.

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] A. Berry, B. Kaba, M. Nadif, E. SanJuan, and A. Sigayret, "Classification et désarticulation de graphes de termes," JADT 2004, Louvain-la-Neuve, Belgium, 2004, pp. 149-160.

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

[16] C. Weaver, "Building highly-coordinated visualizations in Improvise," Proceedings of the IEEE Symposium on Information Visualization, Austin, TX, 2004, pp. 159 – 166.

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

[18] V. N. Vapnik, *The Nature of Statistical Learning Theory*: Springer, 1995.

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