Visual Exploration of Text Collections

VinhTuan Thai

Submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

Advisor	: **Dr. Siegfried Handschuh**
Internal Examiner	: **Prof. Dr. Stefan Decker**
External Examiner	: **Prof. Dr. Enrico Motta**
Examination Chair	: **Dr. John Breslin**

Digital Enterprise Research Institute (DERI) - National University of Ireland, Galway

2012

The studies presented in this thesis were performed at the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway (NUI Galway). The research was financially supported by the Science Foundation Ireland (SFI) under grants No. SFI/02/CE1/I131 (Líon) and SFI/08/CE/I1380 (Líon-2).

Abstract

Despite many technological advances, the information overload problem still prevails in many application areas. It is challenging for users who are inundated with data to explore different facets of a complex information space to extract and put several pieces of facts together into a big picture that allows them to see various aspects of the data. Nevertheless, the availability of data should be embraced, not considered a threat for individuals and businesses alike. As a substantial amount of invaluable information to be explored resides within unstructured text data, there is a need to support users in visual exploration of text collections to obtain useful understandings that can be turned into worthwhile results. In this dissertation, we present our contributions in this area.

We propose an approach to support users in exploring collections of text documents based on their interests and knowledge, which are represented by entities within an ontology. This ontology is used to drive the exploration and can be enriched with newly discovered entities matching users' interests in the process. Coordinated multiple views are used to visualize various aspects of text collections in relation to the set of entities of interest to users.

To support faceted filtering of a large number of documents, we show how a multi-dimensional visualization can be employed as an alternative to the traditional linear listing of focus items. In this visualization, visual abstraction based on a combination of a conceptual structure and the structural equivalence of documents can be simultaneously used to deal with a large number of items. Furthermore, the approach also enables visual ordering based on the importance of facet values to support prioritized, cross-facet comparisons of focus items.

We also report on an approach to support users' comprehension of the distribution of entities within a document based on the classic TileBars paradigm. Our approach employs a

simplified version of a matrix reordering technique, which is based on the barycenter heuristic for bigraph edge crossing minimization, to reorder elements of TileBars-based Entities Distribution Views to tackle the visual complexity problem. The resulting reordered views enable users to quickly and easily identify which entities appear in the beginning, the end, or throughout a document.

Lastly, our work is also concerned with visual concordance analysis, which supports users in understanding how terms are used within a document by investigating their usage contexts. To abstract away the textual details and yet retain the core facets of a term's contexts for visualization, we employ a statistical topic modeling method to group together words that are thematically related. These groups are used to visualize the gist of a term's usage contexts in a visualization called Context Stamp.

Acknowledgments

This thesis would not be completed without the help of many people. I am thankful to Prof. Stefan Decker and Prof. Manfred Hauswirth for giving me the chance to pursue my studies in an intellectually engaging environment. I am deeply indebted to my advisor, Dr. Siegfried Handschuh, for all the support, guidance and encouragement. He introduced me to the information visualization field and gave me invaluable pointers to jump start my research journey. I am thankful to the examiners, Prof. Enrico Motta and Prof. Stefan Decker, for giving me insightful recommendations to improve many aspects of this thesis.

I would also like to thank Dr. Anthony Jameson and Dr. David Huynh for being generous with the feedback and advices on my work. Their words of wisdom certainly broadened my horizons. I also learned a lot of organization skills from Dr. Tom Heath while we ran a couple of workshops together.

I am fortunate enough to have an opportunity to work with supportive and friendly colleagues within the SmILE group, in no particular order: Alexander, Brian, Knud, Fergal, Tudor, Laura, Pierre, Ismael, Charlie, Vit, Simon, Pradeep, Keith, Judie, and Jeremy. Jodi and Sabrina deserve many thanks for all the research discussions as well as their interest in and useful feedback on my work. I very much appreciate the contributions that Pierre-Yves made to the IVEA application during his internship.

Coming to Ireland without knowing anyone here, I am grateful for the many friends who made me feel so welcomed: Fergal, Gearoid, Edward, Aidan. I learned so much from them in many ways.

Lastly, many wholehearted thanks go to my parents and my sister. Though being far away, they have always believed in me and given me the endless support to get through this journey.

Contents

I Prelude **1**

1 Introduction **2**

1.1 Motivation 2

1.2 Approach 5

1.3 Research Questions 6

1.4 Thesis Structure 7

II Background **10**

2 Information Visualization and Visual Analytics **11**

2.1 Information Visualization 12

2.1.1 Overview 12

2.1.2 A Task by Data Type Taxonomy of Visualization Techniques 14

2.1.3 Reference Models for Visualization 16

2.1.4 Visual Information Seeking Process 23

2.2 Visual Analytics 24

2.2.1 Overview 24

2.2.2 Analytical Reasoning 28

2.2.3 A Visual Analytics Process Model 30

2.2.4 Visual Analytics versus Data Visualization 32

2.3 Summary 34

3 Visual Exploration of Text Collections **35**

3.1 Visual Analysis of Text Collections via Dimensionality Reduction Methods 37

3.2 Knowledge-based Visual Text Analysis 50

3.3 Visual Filtering on Text Collections with Query Terms and Facets 54

3.4 Visual Analysis of Document Structures and Term Distributions 65

3.5 Visual Concordance Analysis 76

3.6 Visual Analysis of Trends in Text 86

3.7 Visual Analysis of Text Streams 92

3.8 Other Notable Approaches 96

3.9 Summary 101

III Core **102**

4 Ontology-based Visual Exploration of Text Collections **103**

4.1 Proposed Approach 106

4.1.1 Visual Interface 108

4.1.2 Interactions, Manipulations and Coordinations 111

4.1.3 Implementation 115

4.2 Formative Evaluation 116

4.3 Summary 119

5 Visual Abstraction and Ordering in Faceted Browsing of Text Collections **121**

5.1 Faceted Browsing of Text Collections 122

5.2 A Matrix-based Visualization for Faceted Filtering of Text Collections . . . 126

5.3 Visual Abstraction of Documents 127

5.3.1 Semantic Zooming 127

5.3.2 Documents Grouping 128

5.4 Visual, Interactive Ordering based on Facet Values 131

5.5 New IVEA Visual Interface 132

5.6 Evaluation 134

5.6.1 Method 135

5.6.2 Results and Discussion 142
5.7 Related Work 149
5.8 Summary 151

6 Reordering TileBars-based Entities Distribution Views with the Barycenter Heuristic 152
6.1 TileBars-based Entities Distribution Views 153
6.2 Re-ordering Entities Distribution Views with the Barycenter Heuristic . . . 155
6.3 Evaluation 158
6.3.1 Method 158
6.3.2 Results 160
6.4 Related Work 162
6.5 Discussion 163
6.6 Summary 164

7 Topic-based Content Abstraction for Visual Concordance Analysis 165
7.1 Visual Concordance Analysis 165
7.2 Related Work 166
7.3 Proposed Approach 169
7.3.1 Text Analysis 169
7.3.2 Visualization Design 175
7.4 Evaluation 184
7.4.1 Text Analysis 184
7.4.2 Visualization 184
7.4.3 Discussion 196
7.5 Summary 200

IV Summary 201

8 Summary and Outlook 202
8.1 Summary 202

8.2 Outlook 205
8.3 Enabling Networked Knowledge 208

Bibliography **211**

List of Figures

2.1 Card et al.'s reference model for visualization [12] 16
2.2 Bertin's Retinal Properties [6, 91] 19
2.3 The Data State Reference Model [18] 22
2.4 The integration of different scientific disciplines in visual analytics [70]. . . 26
2.5 Keim et al.'s Visual Analytics Process Model [72] 30
2.6 Effectiveness of analysis vs. degree of interaction [73] 32

3.1 An overview of a semantic network of terms within a selected dimension (fourth) in the 2006 VAST contest data set in Storylines [161]. 40
3.2 Multidimensional Scaling using the Sammon Mapping technique [112]. The numbers in the figure indicate documents that are relevant to queries 1 to 5, the D symbols indicate the remaining documents. 43
3.3 Galaxies and ThemeScape text visualizations of the IN-SPIRE system [155]. 45
3.4 The STARLIGHT text visualization system [108, 107]. 46
3.5 An example of output from the WEBSOM text visualization system [69]. The map contains labels which are examples of the core vocabulary of the area in question. The labels provide an overview of the topics in the document collection. Here areas having light colors contain more documents. . . 48
3.6 A SOM-based map display of a document collection [82]. Clicking on a location in the map will result in a list of top 10 documents semantically related to the term representative of that location. 49
3.7 DocCube [90]. 51
3.8 A screenshot of the SWAPit application [116]. 52

3.9 A screenshot of the PowerMagpie widget [47]. 53
3.10 A visualization showing how documents are related to the three POIs "*laser*", "*plasma*", and "*fusion*" in VIBE [95]. 55
3.11 The transformation from a Venn diagram into an InfoCrystal representing all possible combinations of query terms into Boolean queries. Given three query terms A, B, and C, the interior icons represent: 1 = (A and (not (B or C), 2 = (A and C and (not B), 3 = (A and B and C), 4 = (A and B and (not C), 5 = (C and (not (A or B)), 6 = (B and C and (not A), 7 = (B and (not (A or C)) [125]. 56
3.12 An InfoCrystal showing, for example, 22 documents that are related to the *(Graphical OR Visual)*, *Information Retrieval*, and *Query Language* concepts but not to the *Human Factors* concept (the triangle glyph to the South East of the central glyph) [125]. 57
3.13 Gist Icons being used to represent patents [28]. Clicking on a term (e.g., "*insulin*") results in red circles and red lines indicating the relatedness of that term with respect to all documents containing it. 58
3.14 A example of a faceted browser on a text collection [80]. 59
3.15 The RelationBrowser++ application [160]. 60
3.16 Elastic Lists for faceted browsing on the Nobel prize winners dataset [129]. 61
3.17 The BrowseRDF application [96]. 62
3.18 Aduna AutoFocus [42]. 63
3.19 A TileBars visualization [56]. 66
3.20 A scrollbar-based term distribution visualization [11]. 67
3.21 A term distribution visualization with Focus+Context views [115]. 68
3.22 A visualization of code written by seven different programmers within source files with Seesoft [35]. 69
3.23 A Compus visualization: a collection view on the left and a zoomed in view on the right [38]. 71
3.24 A visual fingerprint showing the lengths of sentences within a document [74]. 72
3.25 Automated visual analysis of document structure [131]. 73

3.26 A structure visualization of four example bills in the Many Bills application [3]. 74
3.27 A set of Document Cards showing key terms and images from four scientific publications [132]. 75
3.28 A Concordance visualization with "*heart*" being the seed term. 77
3.29 A TextArc visualization [97]. 78
3.30 A Word Tree for the seed term "*if*" in *Romeo and Juliet* [151]. 80
3.31 Phrase Nets for Jane Austen's *Pride and Prejudice*, with the patterns "*X and Y*" and "*X at Y*" being used on the left and the right hand side respectively [142]. 81
3.32 The FeatureLens application used with the State of the Union data set [32]. 82
3.33 A POSVis visualization [148]. 84
3.34 A New York Times visualization of a term's usage contexts within State of the Union speeches. 85
3.35 A example ThemeRiver visualization of news articles [53]. 87
3.36 A visual summary of 10,000 emails produced by the TIARA system [83]. . 88
3.37 A *history flow* visualization of the editing history on the Wikipedia article on "Microsoft" [145]. 89
3.38 The Chromograms application with a timeline view of the editing history of a Wikipedia article [152]. 90
3.39 A WikiDashboard of the editing history of a Wikipedia article [134]. 91
3.40 An iChase visualization of one week activity on the WikiProject "*Louvre Paintings*" [106]. 92
3.41 A Themail visualization showing a user's email exchange over 18 months [143]. 93
3.42 A story flow visualization showing the development of different themes within newswires stories [110]. 94
3.43 An Eddi visualization showing major topical themes and related tweets [4]. 95

3.44 Jigsaw's list view of connections between entities within documents. Selected entities are highlighted in yellow. Entities that are connected to others are highlighted in orange [126]. 97
3.45 A DocuBurst visualization of a science textbook, with the seed term "*idea*" [22]. Highlighted in yellow are branches that contain words having the characters '*pl*'. 99
3.46 An example Wordle [146]. 99
3.47 A Parallel Tag Cloud of written decisions of the US Circuit Courts of Appeal [23]. 100

4.1 Approach Overview 106
4.2 IVEA initial interface. 109
4.3 Interactive Filtering by the instance "*Semantic Desktop*" on the X axis and the class "*pimo:Person*" on the Y axis, and Details-on-demand by clicking on the glyph representing the document "C:\data\Papers\iswc 2007\141.pdf", resulting in the corresponding values being displayed in bar charts within the three tabs of the Document Overview panel, and that document's Entities Distribution View also gets updated accordingly. 113
4.4 Adding new concepts or instances to the PIMO ontology. 114
4.5 The PIMO ontology added with a new concept. 115
4.6 Implementation Overview 115
4.7 Questionnaire responses 118

5.1 The initial matrix visualization for faceted filtering. Here a column corresponds to a document and the height of a cell indicates that relevance of a document to an entity. 126
5.2 Semantic Zooming Bigraphs 128
5.3 Semantic Zooming Example 128
5.4 Document Grouping Bigraphs 129
5.5 Document Grouping Example 130

5.6 Facets Reordering Example. In comparison to Figure 5.3, in this Figure, the two entities "*Investigation*" "*Allegation*" have been moved to the top and the matrix was changed accordingly. 131
5.7 The updated interface of the IVEA prototype. In Panel 1, entities of interest to users are shown in a tree structure. Panel 2 is used for faceted filtering. Panel 3 shows the details of a document being in focus among those presented in the filtering view in Panel 2. Panel 4 displays the distributions of entities within the same document in focus as in Panel 3. 133
5.8 UI1 - A linear listing view of the results set. 136
5.9 UI2 - A linear listing view of the results set, added with visual indicators of which entities are contained in each document. 137
5.10 UI4 - A hybrid version of the matrix-based visualization and the linear listing view. 137
5.11 Mean relevant browse speed by UI and task. 143
5.12 Mean subjective ratings. 144

6.1 The Entities Distribution View on the left displays the distributions of entities in a random order from top to bottom. The barycenter heuristic reordering is applied to this view and results in the reordered view on the right. 154
6.2 The mean values of subjects' efficiency scores in different settings. 160
6.3 User satisfaction ratings 161

7.1 A paragraph with topics assigned to words. Words belonging to the same topic 57, such as "*vital*", "*changing*", "*generates*", "*electric*", "*diesel*", "*ethanol*", are in italic. 174
7.2 ContextBurst - the initial visualization used to display usage contexts of the seed term "*insurance*" within the 2010 State of the Union speech. 176
7.3 ContextBurst - Comparison of usage contexts of the seed term "*insurance*" within the 2000 vs. the 2010 speech. 177

7.4 Small multiple Context Stamps at paragraph level with the seed term "*invest*" in the speech given in 2006. Note that the seed term is not shown in this visualization, only in the search box of the application (see the top part of Figure 7.5). 180
7.5 Context Stamps at document level can be used for comparison of contexts across documents in 2007 vs. 2011. 183
7.6 The fulltext display of the contexts of the seed term "*research*" within the 2005 State of the Union speech. 186
7.7 The word cloud display of the contexts of the seed term "*research*" within the 2003 State of the Union speech at paragraph level. 187
7.8 The word cloud display of the contexts of the seed term "*research*" within the 2003 State of the Union speech at document level. 188
7.9 Mean Usability Ratings on the Three Interfaces 195

8.1 A mock-up of a visualization for faceted filtering that incorporates more semantic relations. 206
8.2 A mock-up of a visualization for an Entities Distribution View that incorporates more semantic relations. 207

List of Tables

2.1 Visual features that can be pre-attentively processed [54]. 19

5.1 Comments on Interfaces' Features 147
5.2 Overall Rankings by Percentage of Participants. 148
5.3 Comments on Interface Comparison 148
5.4 Comments on Usability Issues 149

7.1 Findings from the Text Collection by Type 191
7.2 Helpfulness Rankings 192
7.3 Perceived Learning Curves Rankings 193
7.4 Usability Ratings Statistical Tests Results 195

Part I

Prelude

Chapter 1

Introduction

1.1 Motivation

Advances in information and communication technologies have provided us with a wide variety of means to generate, capture, process and store data from different sources, platforms and devices. The advantage of this is that the more data are available, the more their owners can be informed about a domain or a task at hand. The disadvantage is that data also come with a cost. When individuals and businesses alike receive an ever-increasing influx of data on a daily basis, they need to digest the data in order to extract actionable information from them. As such, the availability of a large amount of raw data usually requires a proportional level of effort to cleanse and transform those, sometimes disparate or conflicting, data into a processable or usable form.

Therefore, while the data exploration process should be a joyous experience, in many cases, it turns out to be a major problem of information overload and anxiety [157], when data owners struggle to keep from drowning in the flood of data. The information overload problem refers to the risk of users being lost in data because the data are [70]:

- irrelevant to the tasks at hand
- not processed properly
- presented in an inappropriate way.

Consequences of information overload might be costly. Time, money, and effort might be wasted and potential scientific or business opportunities might be foregone, simply because we are unable to handle the volume of available data [70]. In fact, many application domains (such as finance, business, etc.) rely extensively on the right information being available at the right time [70], so that decision makers can turn their understanding of data into worthwhile outcomes.

Furthermore, it is known that "*the power of the unaided mind is highly overrated. Without external aids, memory, thought, and reasoning are all constrained.*" [92], and hence there are "*fundamental limits that we are asymptotically approaching*". This leads to a kind of "*information glut*", i.e. we have more data at hand than we can process [139]. In addition, exploring data collections becomes increasingly hard as the volume grows [120]. While the amount of data that can be generated and gathered keeps growing, basic human skills and cognitive capabilities do not change significantly over time [139]. As such, it is important that there are external aids to enhance our processing capabilities. In devising external aids, we need to take into account the type of data that we have to deal with. In this thesis, we focus on the unstructured textual data type. This data type is challenging to handle and usually requires various levels of processing before relevant information can be extracted.

As text-based data analysis tasks are known to be cognitively very challenging and taxing on a person's memory, deduction, reasoning, and general analytic capabilities [126], there are various research fields that aim to provide support to users in dealing with textual data. Notably, research efforts within the Information Retrieval area help users in searching for documents that meet their information needs from a large collection of text documents. However, in many cases, apart from the requirement to retrieve documents that are relevant to certain query keywords, users also need to explore and analyze a collection of documents as a whole to gain further understanding. This particular kind of information-seeking activity can be referred to as *exploratory data analysis* or *information analysis* and is commonly carried out in science, intelligence and defense, or business fields [117]. Unlike the information retrieval activity, the exploratory data analysis activity aims to provide users with an overall picture of a text collection as a whole on various dimensions instead of presenting them with the most relevant documents satisfying some search criteria. Furthermore, such an

exploratory data analysis process usually benefits tremendously from data visualization solutions. Those solutions use graphical means to create or discover ideas and understandings by bringing to bear special properties of visual perception to resolve problems [5, 12]. Consequently, visual representations of documents can help people better examine, analyze, and understand them [126]. Depending on the application domains, insights obtained from the visual exploration and analysis of a text collection can enable knowledge workers to identify previously unknown useful information, such as the distribution of topics, clusters of similar documents, trends or linkages between different entities [117].

Given the pervasiveness of textual data, the amount of research work reported in the literature on visual exploration of text collections is vast. In Chapter 3, we will provide an overview of various notable advances in visual text analysis. These works tend to rely on automated text analysis techniques to obtain quantitative findings that can be visualized for interactive exploration by end users. However, most of them do not facilitate users in integrating their interests (e.g., a specific set of entities to focus on) and knowledge (e.g., linguistic variations of each of those entities) into the exploration and analysis of a document collection. Hence, they tend to present findings that are either independent of users' interests or incomplete because of the lack of personal knowledge integration. This is a significant setback when users' spheres of interests and knowledge are of particular importance to the exploratory tasks at hand.

To illustrate the importance of users' interests and knowledge in exploratory tasks, we can consider, for instance, the business analysis domain. In this domain, analysts and investors alike need to explore a large amount of business reports in order to identify, gather and monitor information on companies of interest. In doing so, they can evaluate companies' performance, their competitive positions, and potential investment opportunities. These business reports are, unsurprisingly, information rich as companies are required by law to make disclosures of their financial situation [77]. A significant amount of vital information within these reports is available in free text and not easy to decipher for the casual investors. Such reports are usually considered as a haystack in which "*the valuable needles of telling information can be especially hard to find*" [77]. Furthermore, these reports tend to be written in a rhetorical manner that makes it difficult for investors and analysts to spot the semantic

camouflage and as a result, identify questionable policies or hidden vulnerabilities [77]. In order to extract actionable information, experienced analysts tend to build over time a specific set of entities that are of significant importance to the task, together with their linguistic variations (i.e. different words or phrases that refer to the same thing in their view, such as "*products*","*items*", "*goods*" [135]), and other details, sections to look for within those reports [77]. In a study [99], Pirolli and Card make a similar observation when they analyze the information flow of an analyst's task environment: "*We noted in the interviews and in the analysts' information products that a set of concepts were used repeatedly....We assume that the analyst's sense-making and foraging activities largely proceed by recognizing instances of the categories in the materials scanned (e.g., the entry of a new player into the industry).*" Interested readers can refer to [77] for more examples of the sections and phrases within business reports that are usually of interest to analysts and investors.

The example use case above illustrates a need for the visual exploration process to be aligned with users' interests and knowledge in exploratory tasks within certain domains such as business analysis, investigative analysis, etc., otherwise analysts cannot have a personal viewpoint over the entities that they wish to focus on in order to gain insights. Therefore, in many cases it is necessary to not only perform automated text analysis but also to take into account various important concepts and structures representing users' interest and knowledge. To achieve this, these concepts and structures should be externalized to some formal representation that is user-defined and hence can subsequently be used in combination with text analysis techniques to support the visual exploration process.

1.2 Approach

This dissertation explores a combination of an ontology-based approach and a statistical text mining based approach to support users in visual exploration of text collections. This combination brings about the best of both worlds, whereby a user-defined ontology encapsulating human interests and expertise can be used to drive the exploration process, and a statistical text mining technique enables users to obtain data-driven insights.

Furthermore, we also adopt Shneiderman's well-known visual information-seeking mantra:

"*Overview first, zoom and filter, then details-on-demand*" [120] (described in detail in Section 2.1.4) in designing a suitable visualization application. Taking into consideration the exploratory nature of the task with a focus on integrating users' interests and knowledge and the data source (contents of text documents), the desirable capabilities of a suitable visualization would be:

- Provide an *overview* showing the degrees of relevance of all documents in a collection with respect to users' entities of interests represented by the user-defined ontology.
- Enable users to interactively explore and analyze a text collection at different levels of detail by *zooming in or filtering* the set of documents based on the ontology's concepts and instances.
- Provide a *detailed view on demand* on each selected document, such as the distribution of entities of interest and their usage contexts within each document.

1.3 Research Questions

Based on the requirements discussed in the previous section, this dissertation addresses the following research questions.

1. How can visual interfaces and interactions be designed to support users in exploring text collections based on their interests and knowledge?
2. How can we cater for filtering of a collection of documents in such a way that enables users to easily navigate and explore large results sets?
3. How can we provide support for detailed analysis with respect to the distribution of entities of interest within a document?
4. How can we visualize a term's usage contexts within documents for concordance analysis?

In the next section, we outline how these questions are answered throughout the thesis.

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

In Part II, we provide the background for the research done within this thesis.

- **Chapter 2: Information Visualization and Visual Analytics**

  In this chapter, we go through some of the fundamentals of research within the two areas that our work belongs to, namely Information Visualization and Visual Analytics. This chapter covers basic definitions, process models, principles and distinctions that are helpful for the research and development of information visualization and visual analytics applications.

- **Chapter 3: Visual Exploration of Text Collections**

  In this chapter, we touch on existing works related to the core of this thesis, i.e. those that support visual exploration of text collections. The works discussed in this chapter include the following groups of approaches or applications supporting (1) visual analysis of text collections via dimensionality reduction methods, (2) knowledge-based analysis, (3) visual filtering on text collections, (4) visual analysis of document structures and term distributions, (5) concordance analysis, (6) analysis of trends in text, (7) text streams exploration, and (8) other notable approaches.

In Part III, we present our research contributions.

- **Chapter 4: Ontology-based Visual Exploration of Text Collections**

  In this chapter, we describe an approach toward ontology-based visual exploration of text collections and its realization in a prototype called IVEA. The proposed approach is directed at users who have the need to explore a text collection based on a set of pre-defined entities that are of significant relevance to their exploration goals. The core element in IVEA is a simple personal ontology encapsulating a user's sphere of interest and knowledge. We describe how various statistics about the relationships between documents and entities in the ontology are visually presented by multiple coordinated views to facilitate the exploration. Furthermore, we also discuss a loop of

knowledge utilization and knowledge acquisition/enrichment within IVEA, whereby a personal ontology is used to drive the exploration process and users are enabled to enrich this ontology with newly discovered entities, which are taken into account in later use of the application. The work done in IVEA also paved the way for research reported in the following chapters.

- **Chapter 5: Visual Abstraction and Ordering in Faceted Browsing of Text Collections**

  In this chapter, we detail a novel approach toward faceted browsing of a text collection. Our approach employs a matrix-based visualization to graphically depict the correspondences between documents within a text collection and a set of concepts. In order to tackle the visual scalability issue, which arises when this visualization is used on a large collection, we propose a visual abstraction mechanism based on the structural equivalence of documents and a simple way to visually order groups of documents. In this visualization, the semantic zooming, document grouping and facet reordering features can be used simultaneously and hence enable users to deal with a large amount of documents in a results set. In addition, they also provide users with a flexibility in terms of choosing the levels of abstraction and priorities of the selected facet values.

- **Chapter 6: Reordering TileBars-based Entities Distribution Views with the Barycenter Heuristic**

  In this chapter, we attempt to tackle the visual complexity issue of TileBars-based representations when employed to show the distribution of entities of interest to users within a document. Here we propose using a simplified version of a matrix reordering technique based on the barycenter heuristic for edge crossing minimization in bigraphs to re-arrange elements of TileBars-based Entities Distribution Views. Such an operation results in reordered Entities Distribution Views that enable users to see which entities are only concentrated in either the beginning or the end of a document, and which entities are distributed across a document.

- **Chapter 7: Topic-based Content Abstraction for Visual Concordance Analysis**

  In this chapter, we focus on a topic-based content abstraction approach for visual concordance analysis to support users in getting a quick understanding of and comparing the usage contexts of a term within different documents. The proposed approach is based on a statistical topic model and we rely on its topic-word distributions to group together words that are thematically related. These topical groups are then represented by Context Stamps, which are Treemap-based visualizations showing the decompositions of topics within term-bearing paragraphs at different levels of details. We discuss how the Context Stamp visualization enables users to understand how a term is used as well as to compare and contrast its usage contexts in one document versus another.

In Part IV, we conclude the thesis.

- **Chapter 8: Summary and Future Work**

  In this chapter, we summarize the contributions reported within this thesis and complete it with an outline of potential avenues for further research.

Part II

Background

Chapter 2

Information Visualization and Visual Analytics

As discussed in the previous chapter, the ease with which data can be gathered or generated has resulted in an ever increasing information glut. However, the availability of vast amounts of data is both a challenge and an opportunity. The more data that are at hand, the more informed decisions can be made. In this context, **visualization** is, among others, a research field that contributes solutions to tackle the information overload problem, by tapping into the vast visual processing capabilities of human brains. Furthermore, data visualization research is well-connected with many other disciplines, such as statistics, psychology, human-computer interaction, natural language processing, information retrieval, data mining, etc. to enable users to see, understand, and interact with various kinds of data to get answers to their task-related questions. Visualization plays a significant role in facilitating understanding of available data, as one of its most noteworthy benefits is the amount of information that can be quickly interpreted if it is presented well [150]. Furthermore, as data and tasks become more complicated, there is a need to make sense of complex, conflicting, and dynamic information. This need has provided the impetus for new visual analytics tools and technologies that combine the strengths of visualization with powerful data analysis and innovative interaction techniques [139].

In this chapter, we provide an overview of the information visualization and visual analytics fields. The definitions, processes, and models noted here serve to facilitate understanding

and analysis of various works in the literature that we discuss in the next chapter on visual exploration of text collections.

2.1 Information Visualization

2.1.1 Overview

As data availability is both an invaluable source of knowledge and a challenge in terms of obtaining such knowledge, there is a need for applications that can support comprehension of such data. Most comprehension relies on, among others, our ability to visualize, and this ability is known to be "*almost synonymous with understanding*" [24].

The field of data visualization is, in general, concerned with the research and development of graphical representations of data. Visualization is defined as "***the use of computer-supported, interactive, visual representations of data to amplify cognition***", and there are six major ways in which visualization can amplify cognition [12]:

- By increasing the memory and processing resources available to users.
- By reducing the search for information.
- By using visual representations to enhance the detection of patterns.
- By enabling perceptual inference operations.
- By using perceptual attention mechanism for monitoring.
- By encoding information in a manipulable medium.

The fact that our cognition can be amplified by visualization is important, as while the amount of data to be processed can increase exponentially over time, our working memory and cognitive capabilities do not. It is also worth noting that, similar to computation, whose purpose is insight, not numbers [52], the purpose of visualization is insight, not pictures [12]. Insights obtained from data can play a pivotal role in the discovery, decision making, and explanation processes [12]. To enable users to obtain insights from data, graphical means

are used to bring a large number of data objects into view, revealing patterns and relationships that would otherwise be hard to detect.

As visualization is concerned with different data types, pioneering researchers have proposed a classification scheme to divide works within this field into *scientific visualization* and *information visualization* [12]. This distinction is based on the type of data: *physical* and *non-physical* data. Physical data are inherently geometrical and their visual representations tend to be based on real-world objects, such as the earth, the human body, molecules, terrains, fluid flow around a surface, etc. [12]. Visualizations that show abstractions of physical data based on 3D representations of real-world objects are commonly referred to as *scientific visualizations* [12]. The key goal in scientific visualization is to mimic these objects "*as faithfully as computationally feasible*" [139], and to make "*atomic, cosmic and common three-dimensional phenomena*" visible and comprehensible [120]. Scientific visualizations are hence regarded as tools to enable scientists to handle large collections of physical data and to enhance their ability to identify phenomena within the data [89]. Meanwhile, non-physical or abstract data, such as financial and business data, or collections of documents, do not have any known spatial representations and hence require mappings to visual representations that are suitable for human interpretations. The second branch of visualization, called *information visualization*, refers to "*visualization applied to abstract data*" [12]. Research and development within the information visualization area focus on the design of visualizations that can help users formulate hypotheses, or reveal unexpected properties such as patterns, clusters, gaps, outliers, etc. from huge amounts of abstract data [12]. The widespread use of information visualization in many application domains and mainstream media outlets, such as the New York Times, etc. has indeed proved the prediction that "*information visualization will pass out of the realm of an exotic research specialty and into the mainstream of user interface and application design*" [12].

It is important to note that even though such a classification of visualization into two separate subfields of scientific visualization and information visualization has commonly been used within the visualization research community, it is not without limitations. For instance, in [105] many researchers in the field acknowledge that this simple, widely-known classification scheme is vague and might be problematic and confusing in certain cases. For example,

many data sets are scientific in nature and yet not physically based and hence not inherently geometrical, such as experimental results in chemistry or biology. Notably, Munzner puts it succinctly: "*The current names are rather unfortunate accidents of history: scientific visualization isn't uninformative, and information visualization isn't unscientific*" [105]. The distinction between the two directions of visualization research is thus eventually determined by "*whether the spatialization is given or chosen*" [105]. Despite its limitations, this simple classification scheme still continues to be used in the field, most visibly in the separations of research works presented at VisWeek[1], which is a well-respected forum for advances in visualization, into scientific visualization works at the IEEE Visualization conference and information visualization works at the IEEE Information Visualization conference. In sum, it is useful to understand the historical context of the classification scheme that is being used within the visualization field and the semantics of its subfields' names.

In what follows, we will present some of the core aspects of information visualization that play a key role in the design of visualization applications, based on seminal works in the field.

2.1.2 A Task by Data Type Taxonomy of Visualization Techniques

In order to classify different types of existing visualization prototypes (up to 1996), Shneiderman proposed a **Task by Data Type taxonomy** in [120] . This taxonomy focuses on examples of tasks that are performed on visualizations of the seven data types: 1-dimensional, 2-dimensional, 3-dimensional, temporal, multi-dimensional, tree, and network data. A summary of this taxonomy is as follows [120]:

- **1-dimensional data:** This type includes text-based data such as textual documents[2], software source code, etc. Design issues related to visualizations of this data type may be concerned with decisions on font types, font sizes, colors, and how to show an overview of the text, how to filter out uninteresting items. Potential tasks to be performed on visualizations of 1-dimensional data might be identifying relevant sections

[1] `http://visweek.org/` - Last access date: 2 March 2012

[2] The view taken here of documents is that they are sequences of characters, i.e. more presentation-oriented. This is different from the more common content-oriented view of documents as vectors in a vector space model discussed in the next chapter.

containing certain terms or having some structural properties (e.g., section heads) or filtering for lines of code that were updated most recently. It is worth noting that some visualizations of these data may involve transforming from one data type to another, such as from 1-dimensional text data to data in a multi-dimensional space, which can subsequently be reduced to a representation in a 2-dimensional or 3-dimensional space by dimensionality reduction techniques.

- **2-dimensional data:** This type includes planar or map data such as geographic maps, floor plans, or resulting data from dimensionality reduction of higher dimensional data. User tasks involved with 2-dimensional data include finding adjacent items, paths between items, and other common tasks such as counting, filtering, etc.

- **3-dimensional data:** This type of data is usually related to physical objects with volumetric properties such as buildings, human bodies, objects in astronomy, etc. Therefore, 3-dimensional data are mostly dealt with in scientific visualization research and development. User tasks involving 3-dimensional data might include investigating if there are outliers, patterns, anomalies within the data. Well-known challenges associated with visualization of 3-dimensional data are the occlusion of visual objects and the difficulty for users to understand their current position and orientation within the visual space.

- **Temporal data:** Temporal data play a key role in many application domains, such as the financial or medical fields. Timelines are widely used to visualized temporal data. Examples of user tasks with temporal visualizations include identification of peaks and valleys, or correlations among data items.

- **Multi-dimensional data:** Data from relational or statistical databases can be conveniently manipulated as multi-dimensional data. A data tuple with n attributes is treated as a data point in a n-dimensional space. Visualizations of multi-dimensional data can help users in finding patterns, clusters, correlations between pairs of variables, etc.

- **Tree data:** Data of this type include items in which each has a link to its parent, except for the root. User tasks on tree data usually involve studying the structural properties

of the tree representation.

- **Network data:** This type of data represents relationships between a data item and an arbitrary number of other data items. Examples of users tasks on visualizations of network data might be identifying the shortest or least costly paths between two data items.

2.1.3 Reference Models for Visualization

Creating an effective visualization from raw data is inevitably a challenging task. For this reason, there have been various seminal works to formally characterize elements and processes involved in the creation of a visualization. In this section, we report on notable models well-known in the literature. Despite many technological advances in terms of how data are gathered, stored, and processed, an understanding of these models would be helpful for visualization designers and for researchers in building and comparing visualization applications or techniques.

In [12], Card et al. propose a reference model for visualization, as shown in Figure 2.1. The goal of this reference model is to simplify the discussion of information visualization systems and to make it easy to compare and contrast them. In this section, we summarize the important processes and elements involved in this reference model.

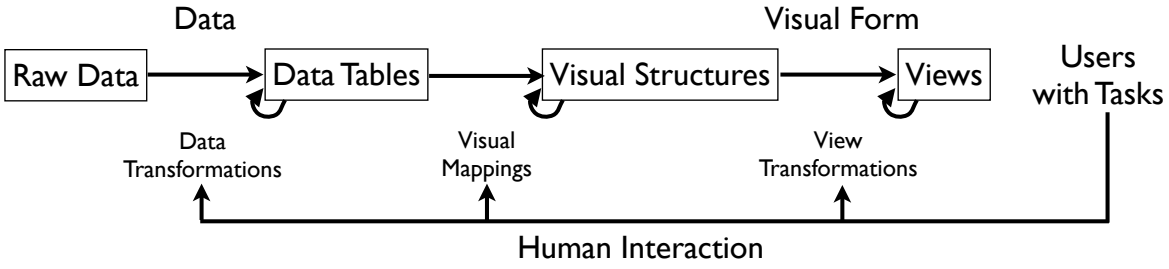


Figure 2.1: Card et al.'s reference model for visualization [12]

- **Data Transformations: Raw Data → Data Tables**

  As illustrated in Figure 2.1, the starting point of building a visualization is to process the raw data. This task is concerned with performing data transformations to convert

data from the original format into one that is suitable for visualization. For instance, raw data can be from spreadsheets, log files, text documents, etc. They usually need to be transformed into data tables representing a set of relations that make it easier to map the data to visual forms.

A data table consists of tuples placed into rows, while variables are placed into columns. Tables of data are also called "*cases by variables arrays*". An advantage of using data tables is that they clearly show the number of variables associated with the data. Data tables with many variables are called *multivariable* data tables. Hence, *multidimensional* visualizations are those designed to encode *multivariable* data tables [12].

Variables can be of the following types:

- *Nominal*, e.g., people names, music genres.
- *Ordinal*, e.g., the four quarters in a year.
- *Quantitative*, e.g., ages. This type of variable can have special subtypes, such quantitative spatial for intrinsically spatial variables that are common in scientific visualization and quantitative geographical for spatial variables representing coordinates.
- *Temporal*, with quantitative temporal variables such as minutes, hours, etc. and ordinal temporal variables such as days of a week.

The transformation from raw data to data tables usually involves a loss or gain of information [12], e.g., cleansing of irrelevant data, fixing errors and missing values, and adding derived values. Data transformations can be classified into the following four types [141]:

- Values $\rightarrow$ Derived Values
- Structure $\rightarrow$ Derived Structure
- Values $\rightarrow$ Derived Structure
- Structure $\rightarrow$ Derived Values

For detailed examples of these types of data transformations, interested readers are referred to [12].

- **Visual Mappings: Data Tables → Visual Structures**

  Even after data transformations, the underlying data structures tend to be neither accessible nor intuitive to users. As such, once raw data have been transformed into tuples in data tables, they need to be mapped onto visual structures, whose role is to augment "*a spatial substrate with marks and graphical properties to encode information*" [12].

  Unlike scientific visualization, in which data correspond directly to physical objects with clear visual structures, in information visualization, data are abstract and hence it is important for visualization designers to choose visual structures that are suitable to convey the nature of the data. Good visual mappings are those that preserve the data, effective mappings are those that are fast to interpret, convey more distinctions or lead to few errors [85].

  To choose appropriate visual structures for a visualization, it is worth taking into consideration some basic aspects of human perception. Previous research has shown that visual information can be processed in two different manners [12]:

  - *Controlled processing*: The processing is detailed, slow, conscious, serial, low capacity and can be inhibited.
  - *Automatic processing*: The processing is fast, high capacity, superficial, parallel, unconscious, and characterized by targets popping out during search.

  Given these two different ways in which visual information is processed, data from tables should be encoded using visual structures which have features that can be automatically processed as much as possible by human perception. In [6], Bertin identifies a set of *retinal* properties, to which the retina of the eye is sensitive. The six basic properties are: size, value, texture, color, orientation and shape, as illustrated in Figure 2.2 (from [91]). The 2-dimensional plane is at the center of the illustration to emphasize the important role of the *position* where these elements can be placed.

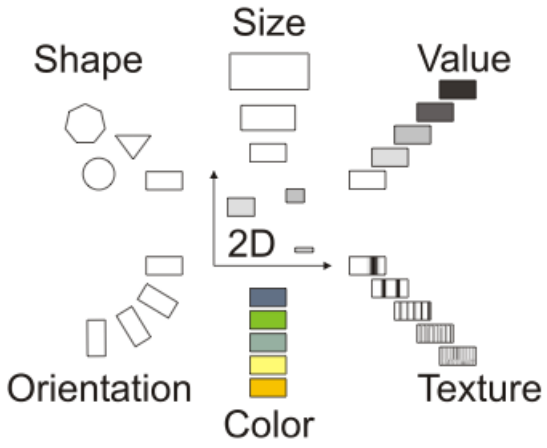


Figure 2.2: Bertin's Retinal Properties [6, 91]

A more recent research work [54] has identified a larger set of visual features that can be pre-attentively (automatically) processed, as shown in Table 2.1.

Number	Terminators	Direction of Motion
Line orientation	Intersection	Binocular Luster
Length	Closure	Stereoscopic Depth
Width	Color	3D Depth Cues
Size	Intensity	Lighting Direction
Curvature	Flicker	

Table 2.1: Visual features that can be pre-attentively processed [54].

It is worth noting that there might be interactions among the visual codings of information. The **Gestalt principles** are a collection of some well-known interactions, which provide insights about basic perceptual phenomena that are valuable for encoding information. These principles are summarized here based on [150] (for visual illustrations, see [150]):

- *Proximity:* Visual elements that are close together are perceived as belonging to a group. Therefore, a simple way to highlight the relationships between elements is to position them in proximity in a display.
- *Similarity:* Visual elements that are similar, e.g., in color, shape, size, texture, orientation, value, tend to be grouped together. Hence, related elements should have resembling looks.

- *Connectedness:* A relationship is usually perceived to exist between visual elements that are connected by lines. Therefore, lines are a powerful visual feature to emphasize relationships in a visualization.
- *Continuity:* Visual elements that are near to each other tend to be grouped together if they are connected by straight or smoothly curved lines.
- *Symmetry:* When two visual symmetrical elements are placed close together, they tend to be perceived as forming a new element as a whole. In addition, symmetry is helpful to design visualizations supporting comparison of visual elements, by arranging them using vertical symmetry.
- *Closure:* A close contour is usually perceived as an object and also there is a tendency to visually close contours with gaps in them. Contours are also helpful to indicate segmentations, as the area within a closed contour tends to be considered separated from the area outside of the contour.
- *Relative size:* Smaller components of a pattern are usually perceived as objects while the bigger part tends to be thought of as a background.
- *Common fate:* Visual elements that move in the same direction tend to be grouped together. This principle is therefore helpful to visualize groups of related elements.

Apart from the Gestalt principles above, there are also other basic principles for developing effective visual representations [92, 139]:

- *Appropriateness principle:* Visual representations should provide neither more nor less information than required for the task at hand. More information than necessary might distract users or make tasks more difficult.
- *Naturalness principle:* Visual representations should closely match the information being presented. Artificial visual metaphors that do not match users' cognitive model of the information might hinder understanding.
- *Matching principle:* Visual representations are most effective when they match the task.

The existence of many principles suggests the importance and the challenging nature of selecting the right visual representation for data to support the task at hand [139]. In fact, poorly designed visualizations might result in incorrect critical decisions and great harm [139]. A well-known example of the significant consequences of not choosing the right visual representation is the Space Shuttle Challenger incident. In this incident, a recommendation to launch the space shuttle in cold weather was made. Part of the reasons for this recommendation was related to a diagram showing the O-ring damages in previous temperature conditions. Unfortunately, this diagram did not make it immediately visible that there was a clear pattern for failures at low temperatures, which eventually led to the uninformed recommendation [140]. A full report on this incident is available on the web[3].

- **View Transformations: Visual Structures → Views**

  Once the visual structures are identified, view transformations generate views from these structures by setting graphical parameters. There are three common types of view transformations [12]:

  - *Location probes:* This kind of transformation involves users interacting with a location in a visual structure to reveal additional information. For instance, in a scatterplot, users can click on any dot to see additional information about the corresponding visual element.
  - *Viewpoint controls:* These controls include zooming, panning, clipping the viewpoint, dynamic queries, etc. They are useful transformations in that they enable users to have different choices of points of view over the data being presented or to focus on a particular subset of visual elements by filtering out those that are not of interest.
  - *Distortion:* This transformation is usually used to create a focus+context visualization. In this kind of visualization, only a portion of the visualization is shown in a large size with all the details so that users can focus on, while the rest of the

[3] `http://science.ksc.nasa.gov/shuttle/missions/51-l/docs/rogers-commission/table-of-contents.html`

visualization is shrunk to keep the context.

Apart from the elements and processed discussed above, two opposite directions stand out in the reference model in Figure 2.1. The direction from left to right in the model illustrates a series of (possibly chained) transformations from raw data to users, while the one from right to left indicates the involvement of human interactions to adjust these transformations. There are many possibilities for human interactions, such as direct manipulation of views to re-arrange visual elements, or dynamic queries to filter out irrelevant items.

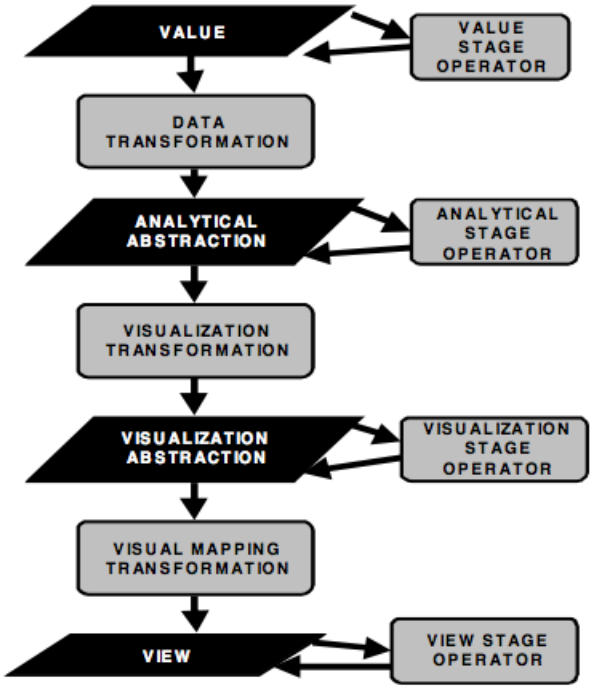


Figure 2.3: The Data State Reference Model [18]

It is also worth mentioning a related reference model for visualization proposed by Chi in [18, 19], called the Data State Reference Model. The motivation of his work is best described as: "*researchers have attempted to construct taxonomies of information visualization techniques by examining the data domains that are compatible with these techniques. This is useful because implementers can quickly identify various techniques that can be applied to their domain of interest. However, these taxonomies do not help the implementers understand how to apply and implement these techniques*" [18]. In this context, Chi proposes the

Data State Reference Model, as shown in Figure 2.3. A brief summary, based on [18], of the stages, processing steps and within stage operators within this model is as follows.

The Value stage corresponds to raw data and the Data Transformation step generates or extracts some form of Analytical Abstraction from the value. Here Analytical Abstraction refers to information about data, and plays the same role as Data Tables in the reference model proposed in [12]. The Visualization Transformation step takes an analytical abstraction and converts it into a visualization abstraction, which is content visualizable on a display. Here the visual abstraction stage is similar to visual structures and the Visual Transformation step resembles the Visual Mappings in Card et al.'s model. The Visual Mapping Transformation step, which corresponds to the View Transformations in Card et al.'s reference model, takes information that is in a visualizable format and shows it in a graphical presentation / view. In addition, within each Data Stage, there is an operator that does not change the data structures. This reference model has been relied upon in the design of well-known visualization toolkits, such as prefuse [61].

2.1.4 Visual Information Seeking Process

While the reference models discussed above are helpful in understanding various elements and processes involved within information visualization systems, it is also important to take into account how users employ such systems to meet their information needs. In this context, Shneiderman proposes the Visual Information Seeking Mantra, which provides a useful guidance to start the design of visualization applications. The mantra focuses on tasks that should be supported in such applications and is stated as follows:
"***Overview first, zoom and filter, then details-on-demand***" [120].

- **Overview:** This task involves getting an overview of the whole collection so that users can have a high-level understanding of the data. As collections tend to contain a large number of data items, it is necessary to make use of additional mechanisms to support users in making sense of the data. Two commonly used strategies are *overview+detail* and *focus+context*. The *overview+detail* strategy involves using an adjoining view as the detail view and there is a movable field-of-view box to adjust the contents of the

detail view. The *focus+context* strategy combines the overview and detail views in one single view, and one or more focus areas are magnified while the rest helps users relate to the entire collection. Fisheye distortion [43], for example, is a well-known *focus+context* technique.

- **Zoom:** Zooming in is helpful for users to focus on a particular portion of the overview display. Therefore, there should be an easy way for users to control the zoom focus and the zoom factor.

- **Filter:** Filtering helps to remove items that are not of current interest. Techniques such as dynamic queries [1] or direct manipulation of data can be used to support users in this task.

- **Details-on-demand:** Once users have zoomed in and filtered out unwanted items, they might need to see the details of a particular item among those left. Therefore, it should be made easy for users to select an item and get all of its details on demand.

While the design of visualization solutions is concerned with many other elements, such as users, tasks, etc., this mantra has been considered useful as a basic guidance in many visualization applications, including ours.

A more in-depth treatment of information visualization as a field is beyond the scope of this section. Interested readers are referred to publications such as [12, 150] for more thorough discussions. In the next section, we move on to describe basic aspects of the Visual Analytics area.

2.2 Visual Analytics

2.2.1 Overview

Advanced data analytics is a critical tool when dealing with such large volumes of data and complex domains that human perceptual and cognitive capabilities as well as visualization technologies cannot scale to the necessary level of complexity. Such an approach of only using data analytics tools is feasible when a problem can be represented in a form that

computer software can handle [71], for instance, market basket analysis from sales data to identify relationships between commodity items based on consumer buying behavior. On the other hand, there are problems that call for complex human judgments and interpretations that are often difficult or impossible to formally specify for computer software to tackle, such as identifying anomalies within the data. This is the class of problems that information visualization can help solve. However, there are other problems that require a combination of contributions from both advanced data analytics and human judgments [71]. For instance, network management systems need to deal with a large amount of data which keep increasing at regular intervals. A visualization cannot display all the available data and hence requires automated data analysis to narrow down the search space. Visual inspection, human judgment, and domain knowledge are then employed to discern network attacks from normal traffic [71]. Visual analytics is the kind of tools that can be used to address such problems, whereby visualization is the medium of a semi-automated analytical process, and humans and machines contribute their distinct capabilities together to produce analytical outcomes [70].

Visual analytics is defined as "***the science of analytical reasoning facilitated by interactive visual interfaces***" [139]. The goal of visual analytics is the creation of tools and techniques to support users in:

- Synthesizing information and deriving insights from data.
- Detecting the expected and discovering the unexpected.
- Providing timely, understandable, and defensible assessments.
- Communicating assessments effectively for action.

From another point of view, visual analytics is considered as follows: "*Visual analytics is the formation of abstract visual metaphors in combination with a human information discourse (usually some form of interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces. It is an outgrowth of the fields of scientific and information visualization but includes technologies from many other fields, including knowledge management, statistical analysis, cognitive*

science, decision science, and others. This marriage of computation, visual representation, and interactive thinking supports intensive analysis. The goal is not only to permit users to detect expected events, such as might be predicted by models, but also to help users discover the unexpected - the surprising anomalies, changes, patterns, and relationships that are then examined and assessed to develop new insight." [24].

It is worth emphasizing that visual analytics research is highly interdisciplinary and integrates a number of research areas, as illustrated in Figure 2.4. As an interdisciplinary field, in many cases visual analytics research requires a team of researchers from different areas to work on specific problems using the expertise from their specific fields [70]. This collaboration across disciplines is likely to bring about two kinds of benefit:

- Collaborative efforts in solving common problems might result in better innovations locally within each discipline, in a more efficient way.
- Integrating relevant results from each of the involved disciplines might lead to significantly improvements in many data analytics application.

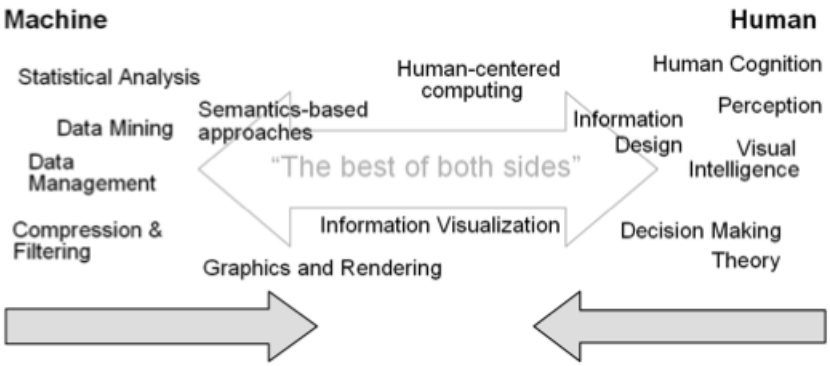


Figure 2.4: The integration of different scientific disciplines in visual analytics [70].

With these characteristics, visual analytics solutions are thus able to potentially turn the information overload and anxiety problems into opportunities. This can be achieved by enabling decision-makers to "*combine their flexibility, creativity, and background knowledge with the enormous storage and processing capacities of todays computers to gain insight into*

complex problems" [73]. As a result, visual analytics is employed in a wide range of application areas including astrophysics, monitoring climate and weather, emergency management, business intelligence, etc.

As the amount of data that visual analytics solutions have to process can be large, one of the most critical challenges in this area is to deal with the scalability issue. The following five major types of scalability issue in visual analytics have been identified [139]:

- *Information scalability:* This issue refers to the capability to extract information from data streams. Dealing with information scalability issues requires methods to filter data, techniques to "*represent data in a multi-resolution manner*", and ways to abstract away the data sets. In addition, there is also a need to address the dynamics of changing data.

- *Visual scalability:* This issue refers to the capability of visual representation and visualization tools to effectively display large data sets, in terms of either the dimension or the number of individual data elements [36]. Addressing this issue will require research into many factors such as: the perception capabilities of the human cognitive system, the visual metaphors used to represent data, the interaction techniques available to users.

- *Display scalability:* Typically, information visualization and visual analytics solutions are aimed at displays on personal computing devices such as desktops and laptops. As data grow and usage scenarios become more diverse, there is a need to take into consideration how those solutions can be flexibly used in large, multiple public displays or small displays such as tablets, smart phones, etc.

- *Human scalability:* As some problem domains are complex, and hence require a collaborative effort from a team of experts, visual analytics solutions might need to accommodate not just a single user usage but rather a collaborative environment. Research and development in visual analytics will need to not only support collaborations between users at the same venue, using a large tabletop display for instance, but also cater to users participating from many different locations.

- *Software scalability:* This important issue refers to the capability of software systems or algorithms to deal with or enable users to interactively manipulate large data sets. Development and maintenance costs are also of concern especially as data keep growing.

In the following sections, we will discuss key aspects of visual analytics including: the analytical reasoning process, the visual analytics process model, and lastly, the differences between visual analytics and information visualization.

2.2.2 Analytical Reasoning

The Analysis Process

Analytical activities are carried out to obtain judgments about important issues based on available data, which might be from different sources. At a high level, the analysis process typically consists of the following phases [139]:

- *Planning:* As a start, analysts need to figure out how to address the issues at hand, what resources to use and how much time should be allocated to each smaller aspect of the main issue.

- *Information gathering:* From the available data, analysts filter for relevant information which can be used as evidence later, become familiar with it and incorporate their own expert knowledge into it.

- *Reasoning:* From their best understanding of the relevant information, analysts generate various candidate explanations or findings, usually by formulating hypotheses. These hypotheses are then validated based on the evidence, together with analysts' knowledge and any assumptions made, to reach conclusions.

- *Documentation:* All judgments made by analysts are documented in reports that summarize the chains of reasoning and the analytical outcomes.

Due to the complexity of available data or the issue at hand, the analysis process is usually done iteratively and collaboratively.

Analytic Discourse

The analysis process described above highlights the different phases involved in reaching judgments from available data using chains of reasoning. Of importance within this process is the analytic discourse, which is "*the technology-mediated dialogue between analysts and their information to produce a judgment about an issue. This discourse is an iterative and evolutionary process by which a path is built from definition of the issue to the assembly of evidence and assumptions to the articulation of judgments.*" [139].

It is therefore important that visual analytics solutions be designed in such a way that they can effectively support the analytic discourse involved. To achieve this goal, Thomas and Cook argue that the following aspects should be considered [139]:

- *Analysis as an iterative and non-linear process:* Apart from gathering information to serve as relevant evidence for making judgments (also called *convergent thinking*), analysts also need to think creatively to make sure that other alternatives are not overlooked (*divergent thinking*). In addition, analysts might also need to consider the broader context of an issue and examine other candidate explanations or data. This activity is also known as *controlled broadening checks*.

- *Actively considering competing hypotheses:* Key to the reasoning activity is the active awareness of competing ideas for the issue being investigated. Therefore, it is important that visual analytics tools facilitate analysts in keeping different candidate explanations active.

- *Enumeration and testing of assumptions:* Making and testing assumptions are important parts in the reasoning process. As such it is necessary to support analysts in expressing them in an explicit representation. Doing so will also help with reviewing the final judgments in a more complete context.

- *Communication of analytical outcomes:* Once analysts have reached conclusions about an issue in question, they will need to articulate and defend their assessments or forecast in ways that meet the needs of the recipients. Visual analytics solutions therefore

should support analysts in conveying judgments, evidence, degree of certainty, and any other alternatives.

An analysis session is hence a dialogue between analysts and the data, whereby visualizations serve as a view into the data and interaction techniques facilitate the dialogue [139]. While most attention has been paid to the design of visual representations, interaction with data does not receive as much focus, despite its important role in visual analytics. Therefore, apart from the three primary uses of human interaction in visualization in the reference model (Figure 2.1) proposed by Card et al. [12], Thomas and Cook add a fourth use for interaction, which is for *human-information discourse*, whose mechanics vanish into the flow of problem solving. Interactions for human-information discourse facilitate processes such as comparing and categorizing data, extracting and recombining data, creating and testing hypotheses, and annotating data [139].

2.2.3 A Visual Analytics Process Model

To provide a high-level overview of the elements and processes involved in a visual analytics solution, Keim et al. propose a Visual Analytics Process Model [72], as shown in Figure 2.5.

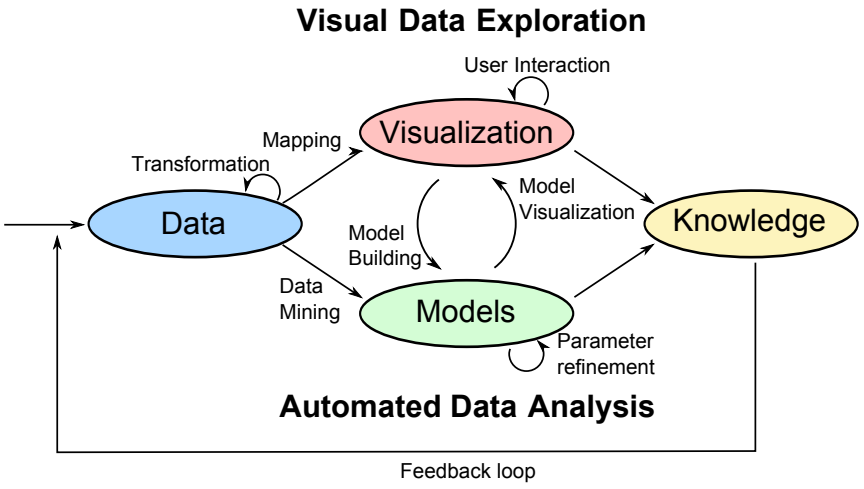


Figure 2.5: Keim et al.'s Visual Analytics Process Model [72]

Within this model, the visual analytics process involves a tight coupling of automated data analysis and visualization methods with human interaction in order to gain insights from data. The first process is data transformation, which prepares data in suitable forms for further use. Typical preprocessing tasks include data cleansing, data normalization, data integration, etc. It is worth noting that raw data are not necessarily the only input into the visual analytics process. Human knowledge, if available, can also be employed given that a suitable knowledge representation is at hand.

Once raw data are transformed into appropriate representations, they can either be directly mapped into a visualization or be processed by data analysis algorithms to build models.

If automated analysis is done first, data mining methods are used to train models from the data. Depending on the application domains, parameter refinement may be necessary to select the most suitable model. For this purpose, visualizations can help by showing visual representations of models, and hence might lead to user interactions with model parameters if necessary. The interactions between visualizations and data analysis methods is considered as "*characteristic for the visual analytics process and leads to a continuous refinement and verification of preliminary results*" [72].

If a direct mapping from data to visualizations is carried out first, analysts can interact with the visualizations to explore their data. It is via these interactions that analysts obtain insights. For instance, they can zoom into areas of interest, or filter out irrelevant items, in order to form hypotheses on the data. Then data analysis methods might be employed to confirm the validity of those hypotheses to obtain insights or knowledge from the data. The knowledge obtained from early iterations of analytics activities can be incorporated into a visual analytics application so that it can make use of that knowledge in subsequent usages of the application.

It is worth emphasizing that while interactions with data models to adjust or refine their parameters to obtain good models are important in many situations, the cases in which they are applicable and whether they are effective depend, to a great extent, on the application domains. Keim et al. identify the following factors as having influences on the proportion of automated data analysis and visualization within a visual analytics solution [73]:

- Users' cognitive capabilities.

- The analysis tasks.
- The available data.

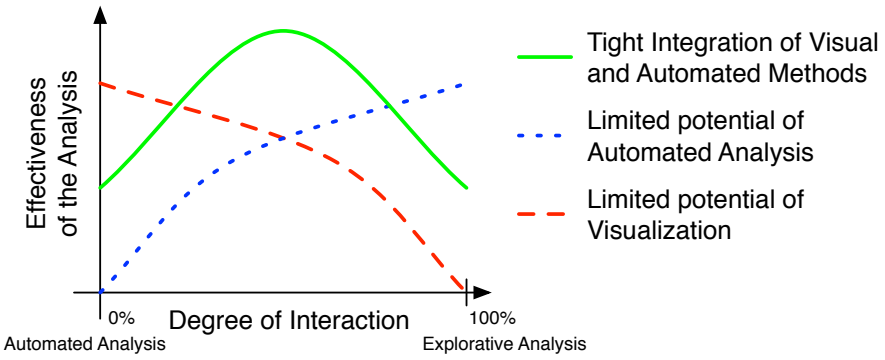


Figure 2.6: Effectiveness of analysis vs. degree of interaction [73]

These factors affect the relationship between the effectiveness of analysis and degree of interaction, as illustrated in Figure 2.6. For problem domains that are well-defined and can be solved by automated data analysis methods (the red dashed line in Figure 2.6), such as the credit card approval task, user involvement might not be necessary and the effectiveness of the analysis might decline as the degree of interaction increases [73]. On the contrary, problems such as the search for Steve Fossett's airplane[4] from a large amount of high-definition satellite images are still better solved by humans using their cognitive capabilitiess (the blue dashed line in Figure 2.6) [71]. In other domains, problems are not well-defined from the outset, hence they might not be solvable separately by neither automated data analysis nor visualization. It is in those cases that a tight integration of the best of both worlds, i.e. the computation capabilities of automated analysis methods with the humans' cognitive capabilities, might be a suitable solution (the solid curve in Figure 2.6) [73].

2.2.4 Visual Analytics versus Data Visualization

In comparison to data visualization, visual analytics as a research area has a much shorter history and as a young field, it builds upon many innovations achieved in data visualization.

[4] http://en.wikipedia.org/wiki/Steve_Fossett

Therefore, it is unsurprising that many find it confusing and hard to distinguish between the two areas. While visualization is highly related to visual analytics, traditional visualization works mainly focus on generating views/ interfaces and interaction techniques for a given data collection, it does not always involve the use of advanced data analysis algorithms [70].

The differences between visual analytics and visualization can best be described as follows [70]: "*Visual analytics is more than just visualization. It can rather be seen as an integral approach to decision-making, combining visualization, human factors and data analysis.* ***The challenge is to identify the best automated algorithm for the analysis task at hand, identify its limits which cannot be further automated, and then develop a tightly integrated solution which adequately integrates the best automated analysis algorithms with appropriate visualization and interaction techniques.***"

As such, visual analytics is perceived to be most different from visualization in its focus on data analysis from the outset and through all iterations of the sense making loop [70]. In addition, it is worth noting that within visual analytics research and development, inventing novel or "*fancy*" visualizations is no longer the main focus. Although some problems necessitate innovative visualizations or interaction techniques, "*most of the times standard techniques work perfectly fine for the problem at hand, because the focus is on finding the most effective solution for the given application problem*" [71]. Keim et al. suggest that the role of visualization within a visual analytics solution is two-fold [71]:

- It provides users with an interface to inspect / understand the automated process and let them adjust it to make the best outcomes from the computation.
- It provides an effective interface to display the results obtained from the automatic analysis process.

Lastly, as visual analytics is still in its early stage, there are many points of view on what it should be. In [13], Chabot, from his experience in the visualization software industry, argues that the following are misperceptions about visual analytics:

- *People adopt visual analytics primarily to see and understand massive data.* While massive data is most likely the input of a visual analytics solution, it is not necessarily so. Visual analytics can be equally useful with small and large data.

- *People adopt visual analytics primarily to see and understand complex data.* Even simple problems/data can be solved quickly with visual analytics, sometimes even with simple visualizations.

- *People adopt visual analytics primarily to see and understand new visual paradigms.* Given the popular utility of visual analytics solutions, there is no need to promulgate new visual paradigms if the existing proven ones are sufficient. Users should not be required to unnecessarily learn new visual paradigms to understand the data if lines, bars, and maps can effectively used to convey the message.

- *People adopt visual analytics primarily to see and understand hidden insights.* While a visual analysis session might result in discoveries of hidden insights, it is not the only way that users can benefit from a visual analytics solution. Most likely visual analytics helps users save time in interactively and quickly performing tasks such as exploring, cleaning, gaining confidence in data, summarizing it, confirming hypotheses and presenting findings.

2.3 Summary

Information visualization and visual analytics are two related types of applications that support users in data exploration. The success of information visualization and visual analytics is characterized by technological advances that enable people to effectively interact with their data to solve their analytical problems. In this chapter, we have discussed the fundamental aspects of both information visualization and visual analytics. In the next chapter, we will cover a particular area of research that is at the core of this thesis, namely visual exploration of text collections.

Chapter 3

Visual Exploration of Text Collections

Text data are pervasive, they come from news articles, business reports, scientific publications, legal documents, patents, customer reviews on e-commerce sites, user-generated contents on social media sites like Twitter and Facebook, etc. Given a large collection of text, the effort required to analyze and consume it could be tantamount to a cognitive challenge for users. For instance, in [155], Wise gives an example of the challenge of dealing with information in textual form within the intelligence analysis domain: "*it is not unreasonable for 30,000 documents to cross the electronic desk of an analyst every week. There is no way that a person could read, retain, and synthesize even one-half of 1% of these.*". The inherent categorical and multidimensional nature of the text data type [59] makes such challenge a significant one for expert analysts and casual users alike. Considering its broad user base, there has always been a need to devise efficient techniques to store, process, retrieve and analyze unstructured text within document collections, as well as interactive tools to support both casual users and expert analysts in navigating and making sense of such a large amount of text data.

In this context, visual text analytics tools and techniques can provide users with the much needed support to overcome the cognitive challenge associated with exploring and understanding text data. More formally, visual text analytics refers to information analysis techniques that enable knowledge discovery from text data via the use of interactive graphical representations [107]. Text visualization is a core element within a visual text analytics solution. While the term "*text visualization*" may refer to techniques used to visualize both

structured and unstructured features of textual data, it is most commonly associated with techniques for displaying the "*semantic characteristics of the free-text components*" of documents [107], which is also the focus of this thesis. Text visualization technologies support our knowledge work in the form of "*computer-assisted knowledge discovery*" to augment our ability to understand and utilize the wealth of information available to us [107]. These technologies enable users of visual text analytics systems to explore multiple dimensions of the document space on an ad hoc basis to derive new knowledge in the form of relationships described by various aspects of documents [107].

In this chapter, we highlight existing works which support visual exploration of text collections. Since in our work we rely on Shneiderman's well-known visual information-seeking mantra: "*Overview first, zoom and filter, then details-on-demand*" [120], we also use it here to relatively organize works in the literature. It is worth noting that many visual text analysis applications do provide all the interactions suggested by the mantra, however, here we use the mantra to loosely separate works in the literature based on their most prominent goal, be it providing an overview of the whole collection, or supporting filtering tasks or helping users investigate documents individually in detail. In what follows, Section 3.1 and Section 3.2 are about approaches to providing an *overview* of the whole text collection. In Section 3.1, we discuss research efforts to provide users with an overview of a document collection via dimensionality reduction methods. In Section 3.2, existing works on providing an overview of a collection based on a knowledge structure are highlighted. In Section 3.3, we move on to visualizations supporting *filtering* tasks on text collections. In Section 3.4 and Section 3.5, we cover works that enable users to see the *details* of individual documents on demand: Section 3.4 reports on tools for visual analysis of document structures and term distributions while Section 3.5 touches upon tools for visual concordance analysis. In addition to the works that can be separated as above, we also provide an overview of tools supporting exploring a more specific type of text collections, i.e. time stamped collections, including: visual trend analysis of text in Section 3.6, visual analysis of text streams in Section 3.7. Even though these two research directions are not within the scope of this dissertation, we are interested in them in future work. Finally, we also discuss other noteworthy approaches in the field in Section 3.8.

3.1 Visual Analysis of Text Collections via Dimensionality Reduction Methods

One of the main challenges in visualization of text is that, compared to the physical space, the document space is much less clearly defined in aspects such as its *dimensionality* (what should be used as dimensions in this space, e.g., words, concepts, etc.), its *measurement* (what values each document should have in each of the chosen dimensions, e.g., word frequencies, binary weights, etc. and how we can measure the distance between two documents in this space) and its *semantic relationship* (e.g., documents that are in close proximity to each other in this space may be considered similar in terms of their contents), and these aspects mostly depend on text processing approaches [82].

Given that the main goal of visual text analytics is to support users in gaining an understanding or insights from a collection of text, research in this field is built upon outcomes from various other inter-related fields, such as statistics, natural language processing, information retrieval, etc. Toward this goal, a significant body of work in visual text analytics deals with generating summary representations of large document collections to illustrate their topical contents and inter-document similarity structures. This particular direction of research is usually referred to as "*semantic mapping*" [107]. For this purpose, a number of computational methods have been developed to characterize and summarize the *semantic structure* of a document space in order to graphically depict the overall conceptual structure of a text collection in a 2D or 3D display [107].

Most text visualization methods for semantic mapping are developed based on vector space models [111] of text collections [107]. In these models, each document is represented by a vector, whose components correspond to weights of terms in the vocabulary, and can be any of (and not limited to) the following options, depending on the design of an application:

- Binary weights: the weight of a term is 1 if it appears in a document, and 0 otherwise.
- Frequency-based weights: the weight of a term is the number of times it appears in a document.
- TF.IDF-based weights: the weight of a term is based on both the number of times it

appears in a document and the number of documents containing that term.

A document collection as a whole can then be represented as a collection of vectors, or more commonly a term-document matrix. For any text collection of a considerable size, the number of unique terms can be rather large, which results in a term-document matrix of significantly high dimensions. In addition, this term-document matrix is often sparse, as only a limited number of unique terms from the vocabulary appear in each document. These unavoidable issues are commonly known as the "*curse of dimensionality*" and the "*empty space phenomenon*" [63]. While this representation is perfectly suitable for information retrieval, it is not so for text visualization [155]. This term-document matrix cannot be visualized as is because it would not be easily comprehensible for users.

In addition, there is no obvious visual representation of the document space, and hence a perceived organization is only created via mappings from the document space to a two-dimensional [82], or in some cases, 3-dimensional display.

These issues necessitate research into dimensionality reduction methods. The key goal of such methods is to map each document from a high dimensional space into a lower dimensional space (typically a 2 or 3 dimensional space for ease of interpretation) in such a way that relationships between documents in the original space are preserved as much as possible in the new space.

In this section, we briefly describe some of the well-known dimensionality reduction techniques and highlight some examples of visualization tools and prototypes that employ them.

Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI) Principal Component Analysis [67] and Latent Semantic Indexing [44] are non-parametric methods that are used in data / text analysis to obtain an approximation of the original term-document matrix in a lower dimensional space. Technically, the goal of both approaches is to identify the most meaningful, linear combination of basis of the original space to re-express a data set, which can help to filter out the noise and reveal the underlying structure of that data set [118].

Given an $m \times n$ matrix X (m is the number of terms used as dimensions, n is the number

of documents), its covariance matrix is considered to identify the noise and redundancy in the data. In this covariance matrix, large values of diagonal terms correspond to interesting structure, and large magnitudes of off-diagonal terms indicate a high redundancy [118]. The Principal Component Analysis approach selects as principal components the ordered set of m orthonormal eigenvectors of the covariance matrix. An application can then choose the first p eigenvectors from this ordered set to be used as the basis of a new, reduced dimensional space. It is expected that the variance along these few principal components (even when p is much less than m) can provide a characterization of the complete data set [118].

Similar to Principal Component Analysis, the Latent Semantic Indexing (LSI) [44] approach also finds the dimensions of greatest variance that cross-cut the term dimensions [107]. In LSI this is done via singular value decomposition (SVD), in which any $n \times m$ matrix A can be decomposed into three matrices: $A = U \sum V^T$, whereby U and V are orthonormal and represents the rows and columns of A respectively and $\sum$ is a diagonal matrix whose elements are a rank-ordered set of singular values. The number of non-zero elements in $\sum$ is the rank of A. To transform the original matrix A to a new matrix in a lower dimensional space, we can truncate the diagonal matrix $\sum$ to retain just the top k singular values, resulting in a matrix $\mathrm{A_k}$ such that: $A_k = U_k \sum_k V_k^T$. The matrix $\mathrm{A_k}$ is the best possible k-dimensional match to the original matrix [107].

Despite their potential in revealing underlying structures in a data set, Principal Component Analysis and Latent Semantic Indexing are not without limitations, such as:

- These linear transform approaches only aim at removing second-order dependencies in the data. However, this is insufficient at revealing all structures in the data if higher order dependencies exist between the variables [118], for instance data that have non-linear structures such as clusters of arbitrary shapes or curved manifolds [98]. This limitation has led to research on non-linear or kernel Principal Component Analysis.

- These approaches are computationally intensive and also require a complete re-calculation when new elements are added into the data set.

In terms of applications, Storylines [161] is an example of visualization applications that enable exploration of document collections as a result of latent semantic analysis. Figure

3.1 shows an overview of a semantic network of terms within the fourth latent semantic dimension in the 2006 VAST contest data set. Each node represents a term and its size represents its contributions to the latent semantic space (c.f. [161] for more detail on the contribution calculation). Links between nodes depict how similar they are in terms of their contributions to the top 100 latent dimensions [161].

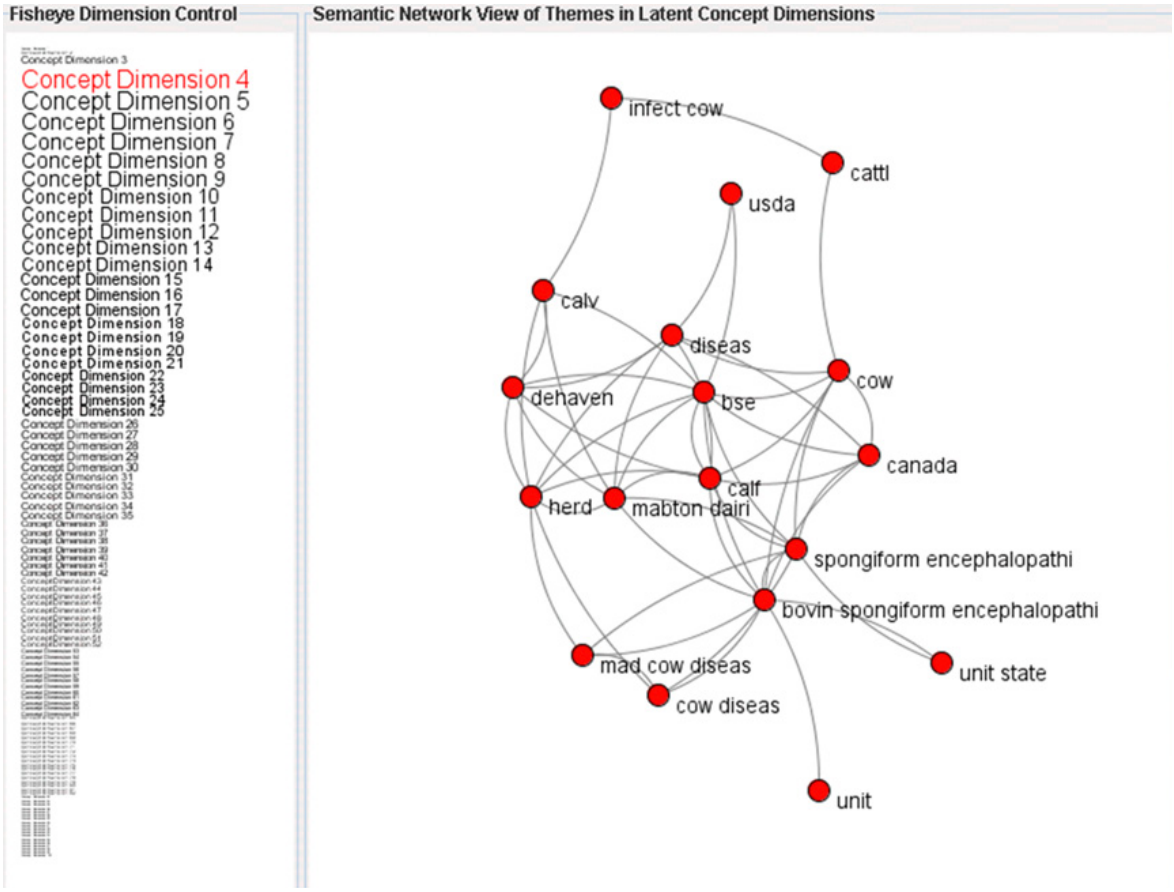


Figure 3.1: An overview of a semantic network of terms within a selected dimension (fourth) in the 2006 VAST contest data set in Storylines [161].

It is worth noting that for different data sets and types of analysis, it is important to determine appropriate parameters when using latent semantic analysis. To support analysts and programmers in understanding the impact and sensitivities of parameters in their models, there exist applications such as LSAView [25]. The approach used in LSAView is to support visual comparisons of the impact of parameter changes on the resulting document clustering outputs. Interested readers are referred to [25] for more detail on the LSAView application.

Multidimensional Scaling (MDS) Multidimensional Scaling is a set of non-linear transformation techniques to project documents (or data in general) from their original high dimensional space to a lower dimensional space in such a way that their salient features are preserved as much as possible. As a result, documents that are conceptually similar are placed close to each other in the resulting *semantic similarity map* [107].

In these techniques, it is important that a measure of pairwise proximity between documents (or data instances) is defined to indicate their dissimilarity. This measure can be obtained by computing the correlation coefficients or Euclidean distances from the vector representations of documents [63]. The Euclidean distance between two documents i and j in an n-dimensional space can be defined as follows (x_{id} is the coordinate of document i in the dimension d):

$$d_{ij} = (\sum_{d=1}^{n} (x_{id} - x_{jd})^2)^{1/2} \tag{3.1}$$

There are two main subclasses of multidimensional scaling techniques: *metric* (or *classical*) and *non-metric* ones. The *metric* techniques are designed to preserve as much as possible the pairwise input distances (or dissimilarities) in the output configuration (the more dissimilar two documents are, the farther apart they will be in the lower dimensional space), while the *non-metric* techniques only attempt to maintain the rank order of the distances [107, 63]. *Metric* multidimensional scaling techniques are more commonly used in text visualization applications [107].

There are two well-known multidimensional scaling techniques: Shepard-Kruskal algorithm [78] and Sammon mapping [112]. Given a set of documents, let ${d_{ij}}^*$ be the distance between document i and j in the original space and d_{ij} is the distance between document i and j in the projected d-dimensional space.

The Shepard-Kruskal algorithm [78] is a non-metric multidimensional scaling technique. It starts with an arbitrary configuration of the targeted d-dimensional space and then iteratively choose a new configuration such that the set of d_{ij} deviates as little as possible from the monotonic ordering of the set of original distance ${d_{ij}}^*$. Kruskal was the first to define a measure called the *stress* function, which measures the lack of fit between d_{ij} and ${d_{ij}}^*$ [78]. There are many variations of the stress function, the simplest version of which can be defined

as:

$$Stress = \left(\sum_{i<j} (d_{ij}{}^{*} - d_{ij})^2\right)^{1/2} \tag{3.2}$$

Given a stress function, the Shepard-Kruskal algorithm minimizes it by using an iterative numerical approach. For a full treatment on the algorithm, interested readers are referred to [78].

Related to the Shepard-Kruskal algorithm is the Sammon mapping [112] approach. This is a metric multidimensional scaling approach in which an error function E is defined to represent how well a configuration of the data points in the projected d-dimensional (d is usually 2 or 3) space fits the original n-dimensional space. E is defined as follows [112]:

$$E = \frac{1}{\sum_{i<j}[d_{ij}{}^{*}]} \sum_{i<j}^{n} \frac{[d_{ij}{}^{*} - d_{ij}]^2}{d_{ij}{}^{*}} \tag{3.3}$$

The Sammon mapping algorithm then proceeds iteratively to find a d-space configuration that minimizes the error E using a steepest descent procedure. For detail of the algorithm, interested readers are referred to Appendix I in [112]. Figure 3.2 is an example of how 188 documents in a 17-dimensional vector space are scaled into a 2-dimensional space. The numbers in the figure indicate documents that are relevant to queries 1 to 5, and the symbol D indicates the remaining documents in the collection [112]. This graphical display is considered as one of the earliest visualizations of text document collections as a result of a dimensionality reduction approach [107].

The Bead system [15, 16] is also one of the fundamental works which use multidimensional scaling techniques to generate semantic maps of document collections in a 3D environment. Bead employs the Force-Directed Placement (also known as gradient descent) approach in which the differences between the distances of objects (such as documents) in the targeted space and in the original space is minimized by spring forces. These forces tend to pull objects that are similar but distant toward each other, and push away those that are dissimilar and close [14]. Bead is worth noting here as the algorithm proposed in [14] reduces the running time of multidimensional scaling techniques from a typical quadratic to a linear one by using stochastic sampling and neighbor sets (c.f. [14] for technical details).

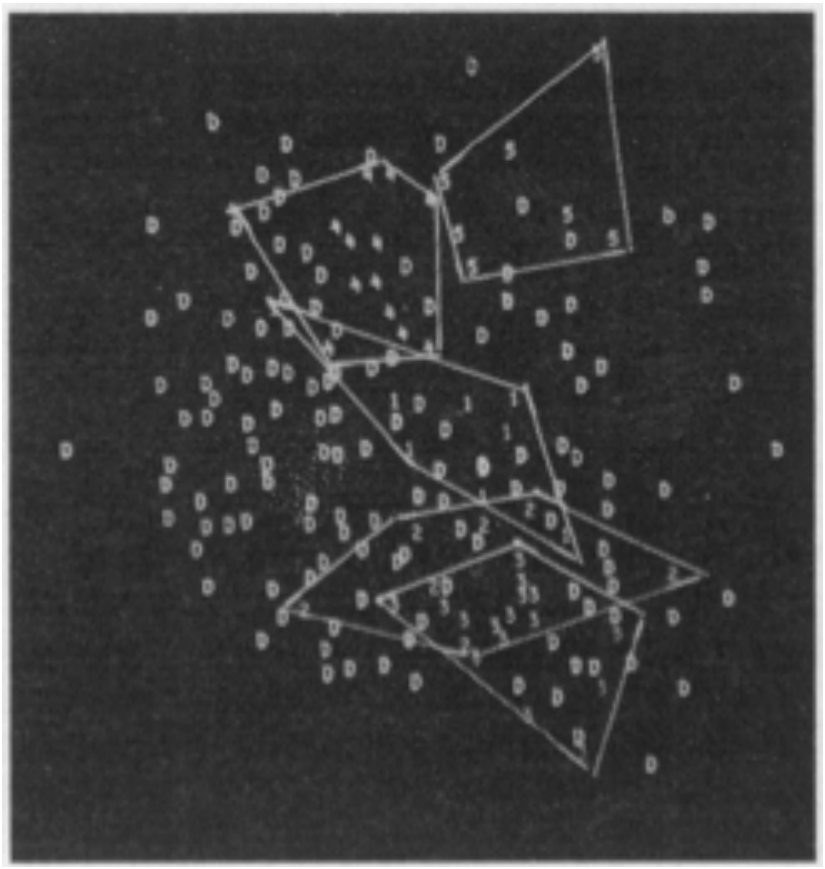


Figure 3.2: Multidimensional Scaling using the Sammon Mapping technique [112]. The numbers in the figure indicate documents that are relevant to queries 1 to 5, the D symbols indicate the remaining documents.

Another well-known example of visual text analytics systems employing a multidimensional scaling technique is the IN-SPIRE system [156]. The text visualizations made available within the IN-SPIRE system are the 2-dimensional Galaxies view and the 3D view called ThemeScape, as shown in Figure 3.3. The system initially used a multidimensional scaling technique, and then anchored least stress, and finally hybrid clustering/principal component analysis for better efficiency in mapping documents to a 2D space. It also used the landscape metaphor in a 3D view to depict the relative prominence of different themes in a text collection [107].

The STARLIGHT system [108], as shown in Figure 3.4, employs a similar hybrid approach to the IN-SPIRE system in that it combines clustering with a multidimensional scaling technique to derive a semantic similarity map of clusters of documents more efficiently. This improved efficiency is obtained because instead of performing multidimensional scaling on all documents in a collection, which is computationally intensive, the STARLIGHT system first clusters documents into a small number of groups of thematically related documents, and then multidimensional scaling is applied to map the centroids of the clusters into a 2D or 3D space [108]. Similarly, documents within each cluster are then mapped to a local coordinate system. This in effect generates semantic similarity maps at multiple levels of granularity [107].

Self-organizing map (SOM) A self-organizing map (or a self-organizing feature map) [75, 76] is an unsupervised-learning neural network that produces a similarity graph of input data. In the case of text documents as input data, each data instance is originally represented by a vector of terms and a learning process automatically maps them onto nodes of a 2D grid according to their mutual similarity. At the beginning of the training process, a set of output nodes are randomly represented by vectors of random weights. The mapping procedure is a recursive regression process which can be summarized as follows [75, 82]:

- Randomly pick an input vector from the data.
- Search for a node on the grid which is closest to the chosen input vector above in the original space. The proximity is usually defined as the Euclidean distance. The closest node is called the winning node.

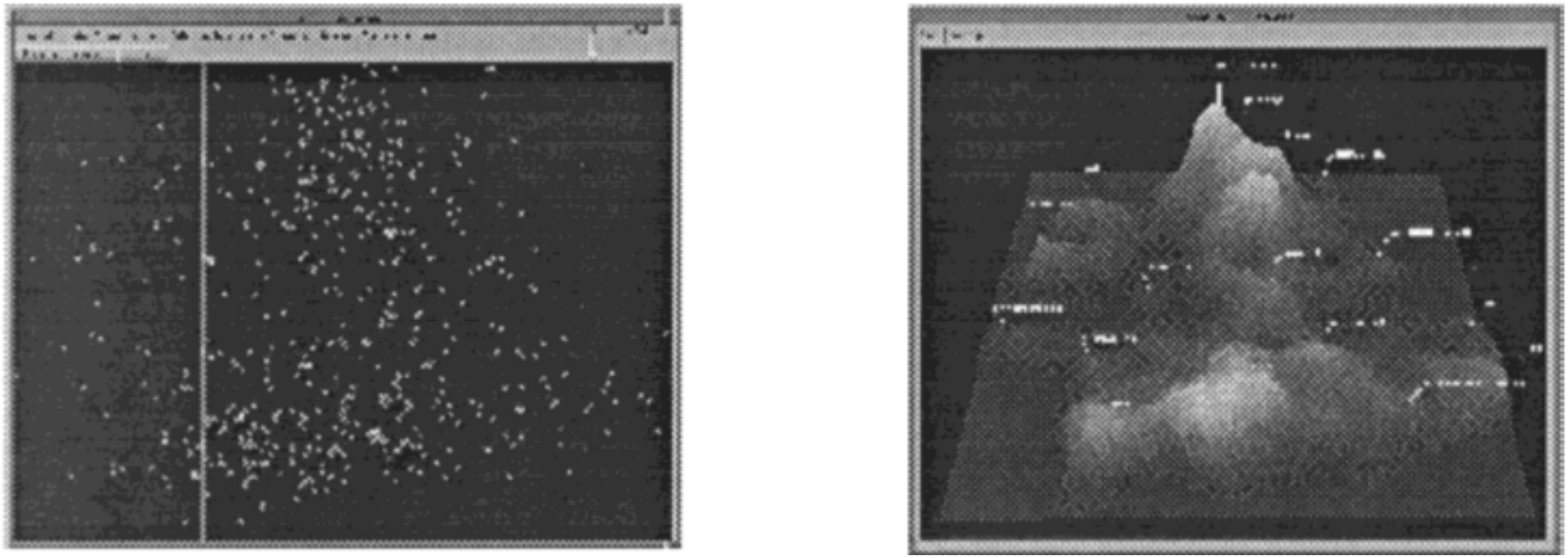


Figure 3.3: Galaxies and ThemeScape text visualizations of the IN-SPIRE system [155].

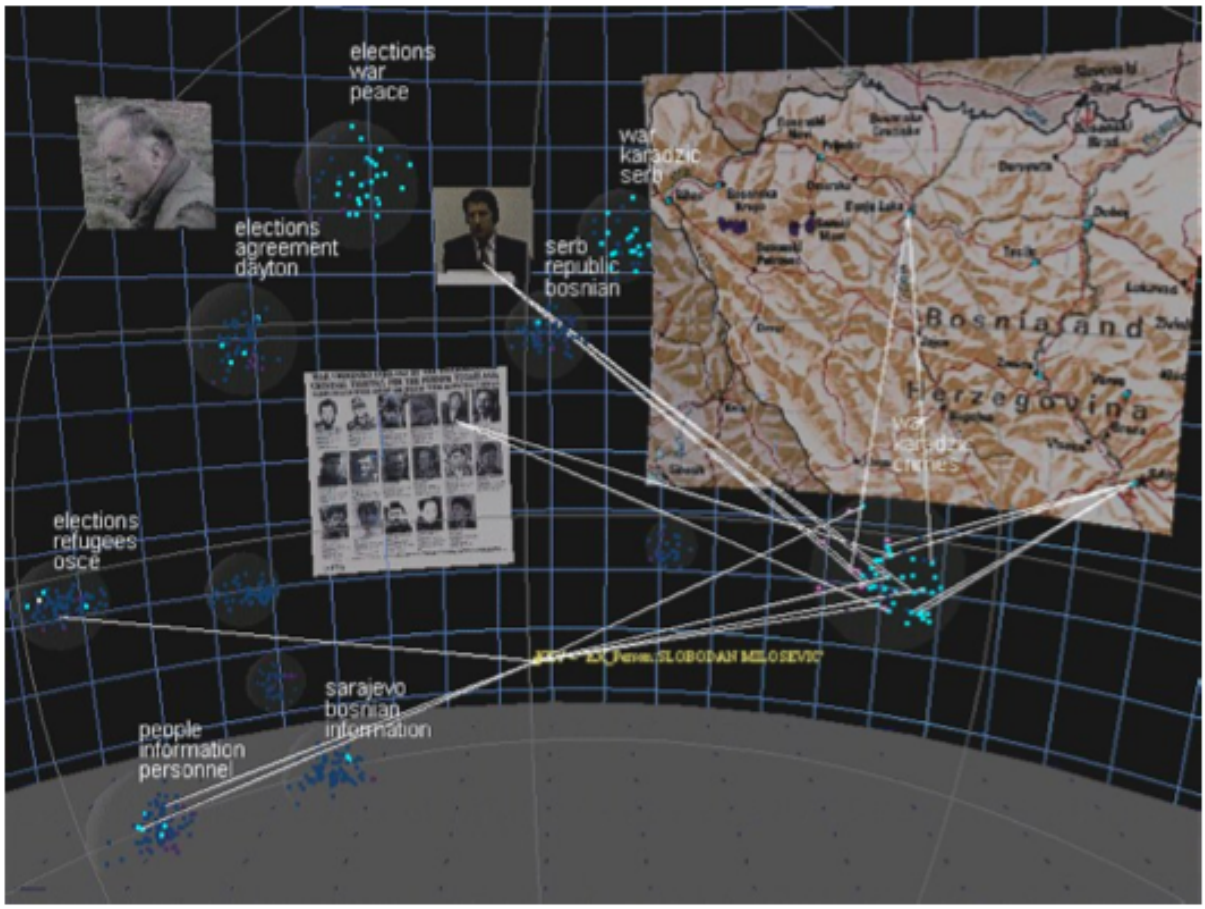


Figure 3.4: The STARLIGHT text visualization system [108, 107].

- Adjust the weights of the winning node such that it is closer to the chosen input vector.
- Adjust the weights of the nodes that are neighbors of the winning node, so that nodes within this neighborhood will have a similar weight patterns.

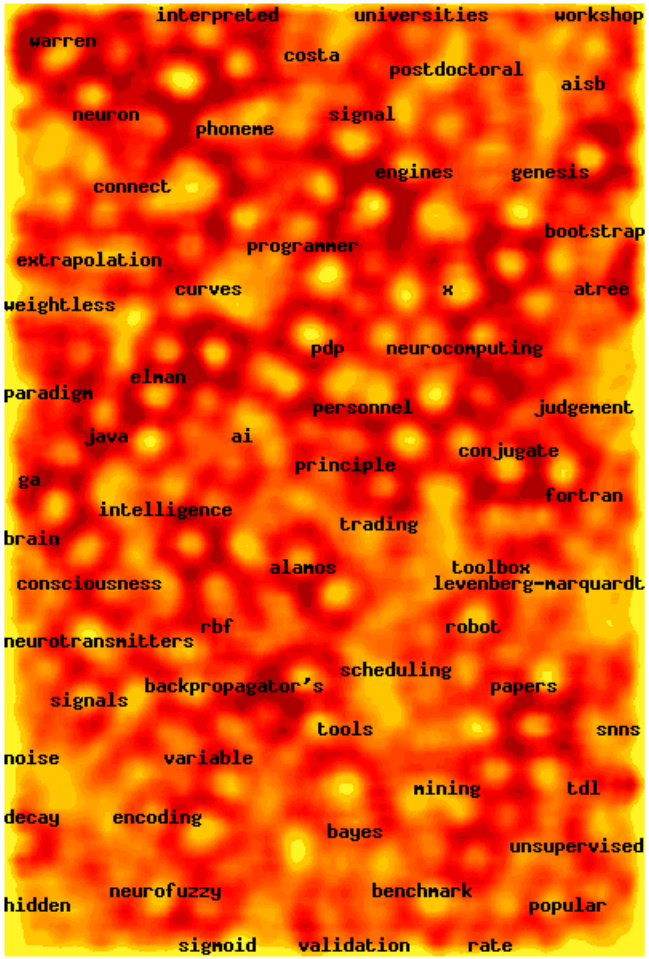
The above process goes through a number of iterations until it converges, i.e. the adjustments approach zero. The neighborhood selection in the last step needs to be defined in a function that shrinks gradually during the process [75]. This recursive learning process eventually turns a grid of nodes with randomly assigned weights into an orderly feature map that reflects the structure of the input data [82].

WEBSOM [69] is an application of the Kohonen self-organizing map algorithm to web data. Figure 3.5 shows a zoomed-in portion of a document map generated by the WEBSOM system[1] from a large document collection. The documents are in the points of the map, and their contents can be accessed by clicking the points visible on the lowest level of the map display.

In [82], Lin proposes an approach based on self-organizing map to support information retrieval tasks. In addition to using a map display, the approach also uses terms as labels for areas on the map which might help to make interpretation easier. Those labels are selected at the end of the training process, whereby the weight of every term is compared with the weight of every output node to select the best term for each node. Nodes that share the same representative term are joined to generate areas on the map [82]. Figure 3.6 is an example of a map display generated by the approach from a collection of documents in the proceedings of the SIGIR conferences from 1990-1993. When users click on a location in the map, a list of top 10 documents semantically related to the term representative of that location is shown [82].

Limitations Even though dimensionality reduction methods are helpful to provide an approximation of data from a multi-dimensional space to a much lower dimensional space, they have certain known limitations including:

[1] `http://websom.hut.fi/websom/`



Automatically generated labels and examples of titles in which the labels have occurred

---

ai - SubSymbolic AI :-O
aisb - AISB and ECCS 1997 - Call for Registrations
alamos - Graduate Research Assistantship in Los Alamos National Laboratory
atree - Atree 3.0 EK mirror sites
backpropagator's - Backpropagator's review, by Donald R. Tveter
bayes - nn and stat decisions for med: Bayes vs neorotic nets
benchmark - Benchmark data for signal processing
bootstrap - Bootstrapping vs. Bayesian
brain - Brain usage Was:Re: Function of sleep
conjugate - Moller's Scaled conjugate gradient (Re: Questions about my new NN program
connect - Direct Connect Problems
consciousness - electromagnetics, brain waves and consciousness
costa - New paper available "The Application of Fuzzy Logic in Automatic Modelling Electromechanical Systems" - P.J. Costa Branco
curves - Predicting Learning Curves (Ph.D. thesis, online)
decay - weight decay
elman - Elman (or Edelman?) nn, references please.
encoding - Question on Encoding Input Data
engines - IC ENGINES & NEURAL NETS
extrapolation - Generalization (Interpolation & Extrapolation)
fortran - Neural Net Fortran Code

Figure 3.5: An example of output from the WEBSOM text visualization system [69]. The map contains labels which are examples of the core vocabulary of the area in question. The labels provide an overview of the topics in the document collection. Here areas having light colors contain more documents.

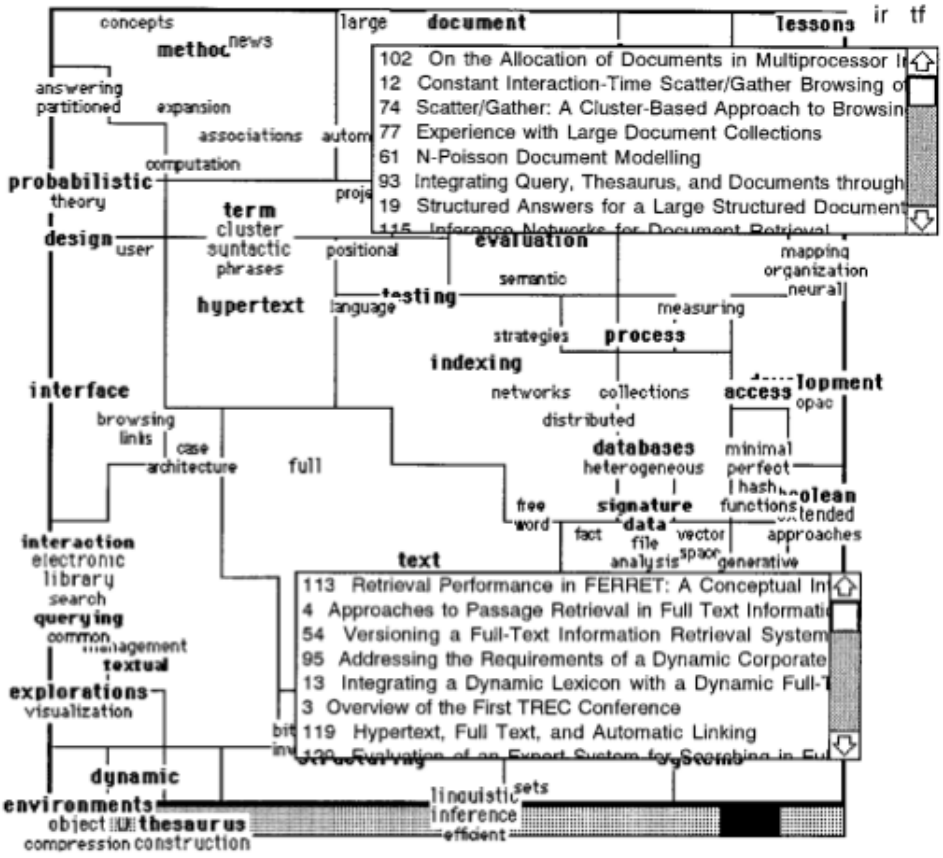


Figure 3.6: A SOM-based map display of a document collection [82]. Clicking on a location in the map will result in a list of top 10 documents semantically related to the term representative of that location.

- Computationally intensive: The computational complexity of techniques such as linear or non-linear Principal Component Analysis or metric Multidimensional Scaling makes it necessary to consider simultaneously dimensionality reduction and scaling [156].

- Difficult for users to comprehend: The semantics of the newly generated spatial dimensions are not easy to define and might hinder or lead to different interpretations [59, 74, 95].

- Significant loss of information: Due to the need to visualize data in a two or three dimensional display, it is usually necessary to eventually use only the first two or three principal components as the dimensions of the new data space. However, this often leads to a significant loss of information [95, 63].

3.2 Knowledge-based Visual Text Analysis

The approaches described in the previous section are based only on the contents of documents. In this section, we discuss applications that can provide an overview of a collection based on a knowledge structure such as a domain ontology.

The DocCube application employs an approach based on OLAP principles to guide users in exploring a text collection [90]. DocCube uses a knowledge structure such as a concept hierarchy or a domain ontology to support the exploration. Given a set of predefined categories, a supervised learning method is used to learn which terms a category's *profile* should contain, and the weights of the associations between documents and categories are based on the number of terms within a category's profile contained in those documents [90]. The DocCube interface employs 2D and 3D scatter plots to show the number of documents belonging to several categories [90]. Figure 3.7 shows how documents are associated with three different categories, each sphere represents a set of documents related to the selected categories and the document list shows all documents in such a sphere. There was no usability study reported on this system. The use of a 3-dimensional space means that DocCube suffers from the well-known occlusion issue and hence hinders navigation, especially when exploring a

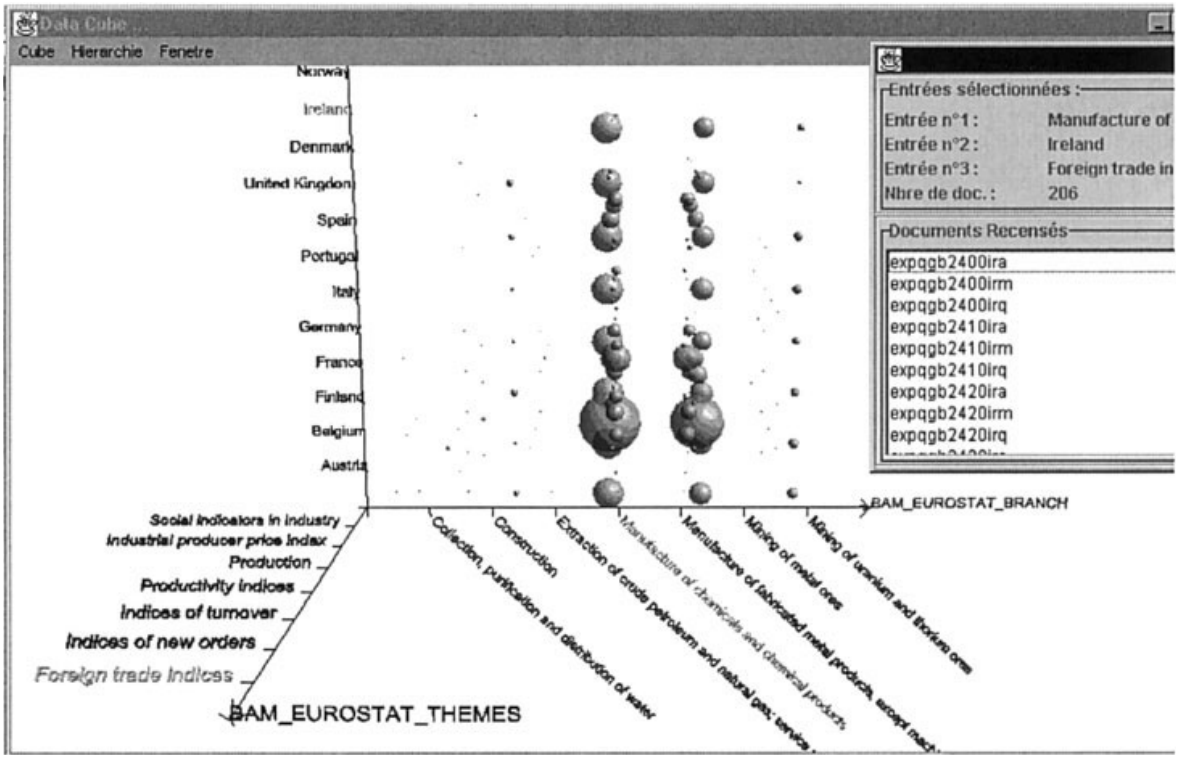


Figure 3.7: DocCube [90].

large text collection. Furthermore, the supervised learning approach requires that training data be available for each domain ontology used for exploration.

In SWAPit [116], a mixed approach toward exploration of a collection of text documents is proposed by using both a multidimensional scaling method and a domain ontology encapsulating an analyst's interest profile. In this approach, documents are the smallest unit of analysis. The SWAPit tool provides multiple coordinated views, as shown in Figure 3.8, in which selections made in one view are highlighted in all other views. A document map, as shown in the upper left hand side of Figure 3.8, is used to visualize the similarity between documents, as a result of a multidimensional scaling method. Each document is represented as a dot on this map and similar documents are positioned close to each other. Selecting documents on the map will result in highlighting of concepts in the ontology tree (below the document map) they are related to, and vice versa, selecting either domain concepts or records from the relational data table will make related documents decorated with colored squares on the map. SWAPit employs a set of visual icons in such a way that documents

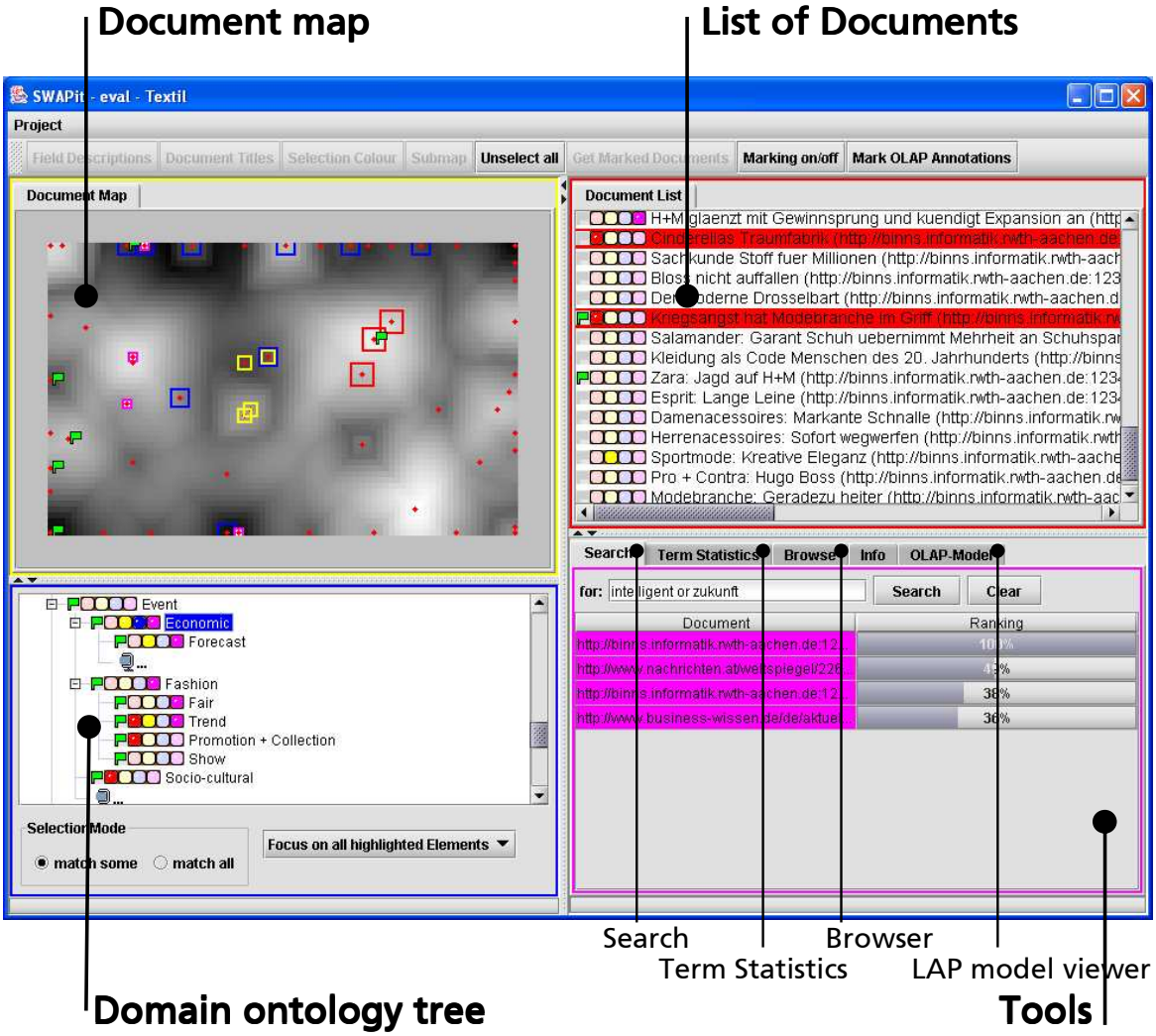


Figure 3.8: A screenshot of the SWAPit application [116].

selected from a particular view are assigned a color. For instance, in Figure 3.8, documents selected from the map will have yellow squares on the map, and the corresponding entries in the relational view will have the subdued yellow icons darken and the ontological classes that these documents are assigned to also have their icons darken with yellow. Other views have other colors to indicate the sources of selection (e.g., selections from the ontology tree correspond to blue icons, etc.). It is unclear from this work how usable this visual encoding scheme is.

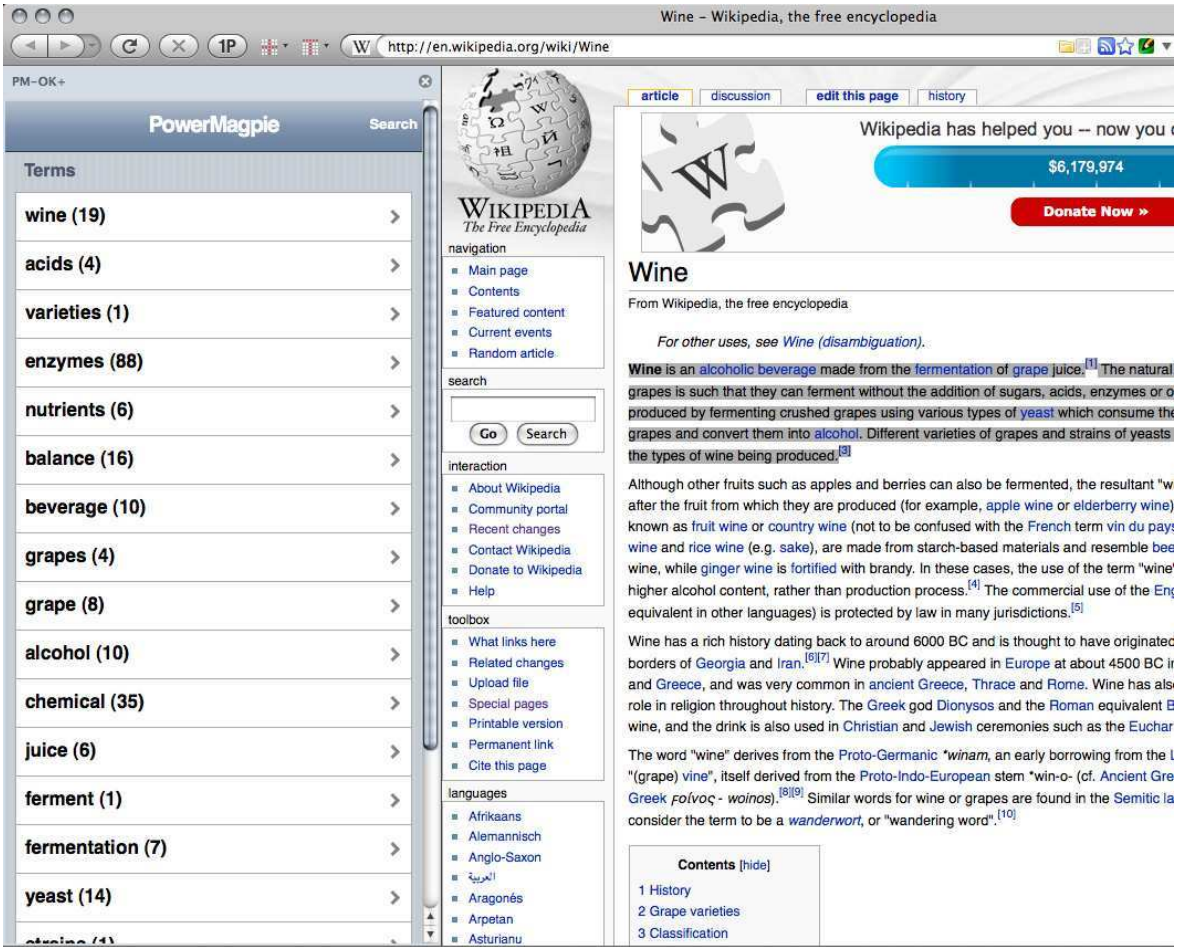


Figure 3.9: A screenshot of the PowerMagpie widget [47].

A semantic web browser called PowerMagpie [47, 48], as shown in Figure 3.9, also relies on ontologies to support browsing of documents on the web. PowerMagpie is different from DocCube and SWAPit in that its focus is on supporting users in linking ontological entities to their mentions in the text of web pages, instead of enabling visual exploration at different

levels of aggregation. Another major difference is that in PowerMagpie, ontologies that are relevant to the current page are searched using a search engine called Watson [27], and presented to users so that they can explore entities from those ontologies within the page. In addition, PowerMagpie also presents users with details of ontologies' concepts and relations on demand.

3.3 Visual Filtering on Text Collections with Query Terms and Facets

The approaches discussed in the previous sections aim at presenting users with an overview of how documents are similar to each other or how they correspond to a knowledge structure. In this section, we move on to highlight some of the works whose focus is to support filtering tasks that help users narrow down a large collection of documents to a smaller subset. Filtering can typically be done via *query terms* or a pre-defined *hierarchy of facets* (*faceted browsing*).

The approaches that support filtering via *query terms* tend to display the relationships between the filtered documents and the subset of the query terms that they contain. For instance, in the case of VIBE [95], it employs a 2-dimensional display to support users in filtering for documents within a collection that are related to their points of interests (POI), each of which consists of a set of keywords relevant to a subject of interest to users. VIBE aims at providing an understanding of the correspondence between these documents and the selected POIs. In VIBE, the positioning of relevant documents on a 2-dimensional display is based on the degrees of overlapping between keywords that users define for the POIs and documents' contents. Figure 3.10 shows documents related to the three POIs "*laser*", "*plasma*", and "*fusion*". Documents closer to any POI are more relevant to it. As intuitive as the VIBE visualization seems, visual scalability is an issue when used with a large document collection. Moreover, while VIBE does not suffer from limitations of dimensionality reduction methods, such as loss of information and difficulty of interpretation of new dimensions, there are cases whereby this representation can be confusing for users, i.e. when there

are four or more POIs. For instance, if there are four POIs called A, B, C, and D placed clockwise at four corners of a rectangle, it would not be possible to tell if the glyphs appearing at the intersection of the lines AC and BD represent documents related to all four POIs, or just to (A & C), or (B & D) [125] (interested readers can refer to [125] for a graphical illustration).

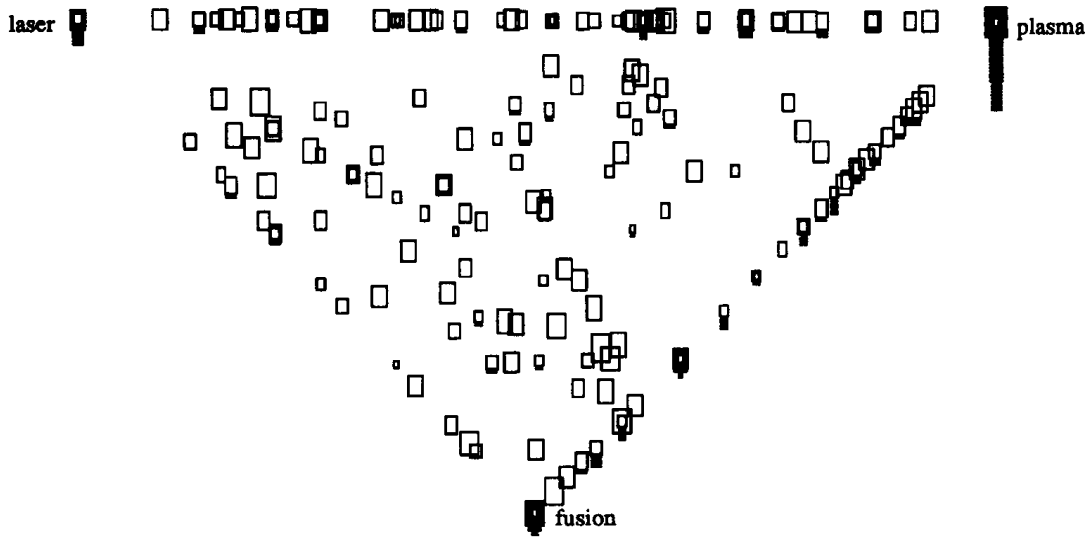


Figure 3.10: A visualization showing how documents are related to the three POIs "*laser*", "*plasma*", and "*fusion*" in VIBE [95].

Similar to VIBE, InfoCrystal [125] also aims to support exploring a collection of documents via filtering based on users' interests. An InfoCrystal display, as shown in Figure 3.11, is derived from a Venn diagram given N query terms used for filtering. In an InfoCrystal, each query term is placed at a corner of a polygon, and each glyph within this polygon is created to represent a group of documents containing one of the $2^N - 1$ possible combinations those query terms. These glyphs are mainly coded by the number of terms matched (e.g., circles for 1, rectangles for 2, triangles for 3 in Figure 3.11), proximity (glyphs closer to a term represent documents that are more related to that term), rank (the closer a glyph is to the center, the more terms its document contains, and the ranks are implicit concentric circles, e.g., in Figure 3.11 documents 1, 5, 7 have rank 1, documents 2,4,6 have rank 2, and document 3 has rank 3) [125]. A setback with this layout is that in the case of glyphs having rank 2 that lie opposite each other on the diagonal line (glyphs representing groups of 84 and

90 documents in Figure 3.12), they need to be shown twice because "*the ranking principle takes precedence over the proximity principle*" [125]. This, however, might cause confusion to users who are not familiar enough with this visualization to understand that those two groups are actually the same. In addition, the large number of different encodings in use might make it challenging for users to comprehend this visualization.

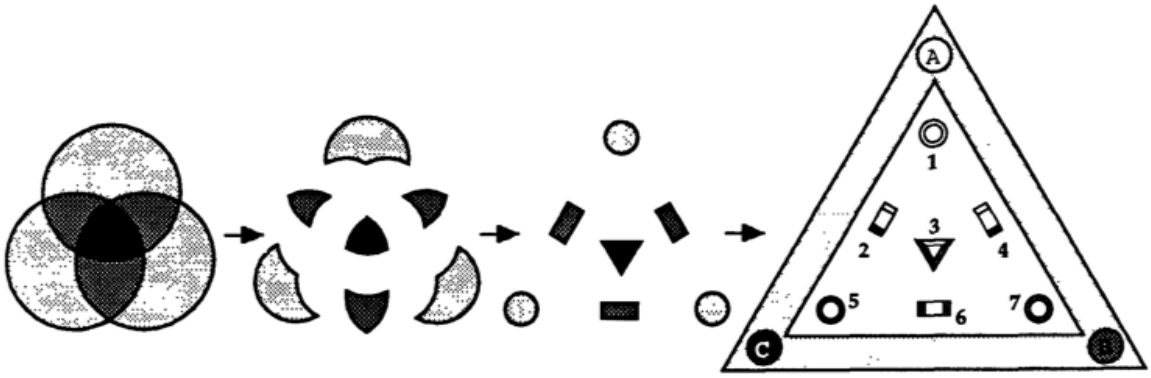


Figure 3.11: The transformation from a Venn diagram into an InfoCrystal representing all possible combinations of query terms into Boolean queries. Given three query terms A, B, and C, the interior icons represent: 1 = (A and (not (B or C), 2 = (A and C and (not B), 3 = (A and B and C), 4 = (A and B and (not C), 5 = (C and (not (A or B)), 6 = (B and C and (not A), 7 = (B and (not (A or C)) [125].

Another noteworthy approach to support filtering tasks is Gist Icons [28], as shown in Figure 3.13. Gist Icons are compact representations of the relatedness between documents and words. For each document, a histogram shows how it is related to a set of words, which are organized in such a way that related words are near to each other in order to form some sort of loosely defined concepts. This histogram, however, is wrapped around into a radial layout. A smoothed contour is then drawn to create a shape which becomes the document's visual profile (the gray shape in Figure 3.13). The peaks in the shape can hence signal a document's relatedness to some concepts while the valleys indicate the concepts it is unrelated to. Gist Icons are designed such that exploration of 50-100 documents is possible at a time. A separate fisheye histogram (on the left hand side of Figure 3.13) shows the average relatedness values of all documents, and when users click on a term, the relatedness values of that term with respect to all documents are visually highlighted by red circles and red lines. Documents with similar profiles are grouped together, and colored green as shown

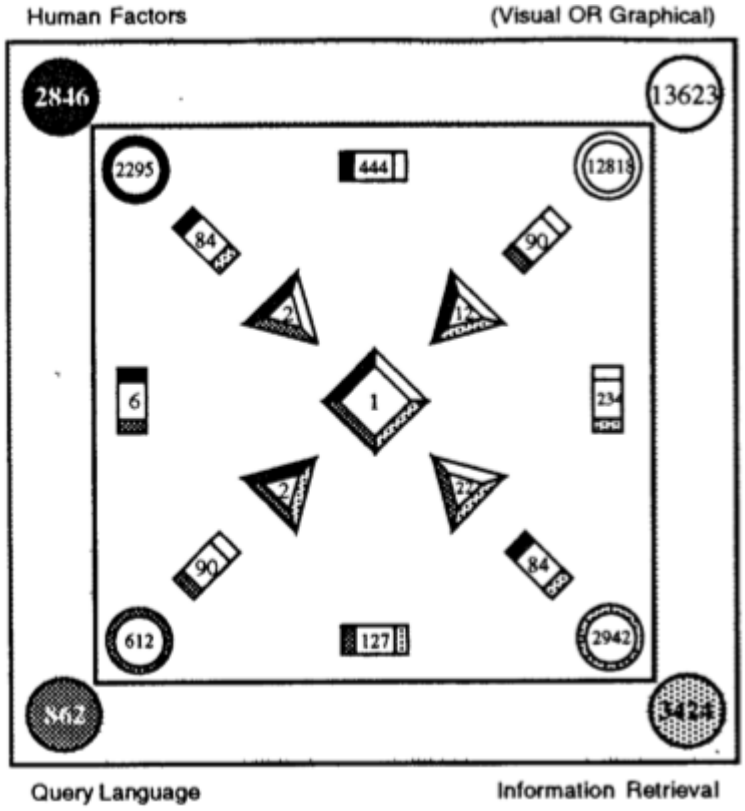


Figure 3.12: An InfoCrystal showing, for example, 22 documents that are related to the *(Graphical OR Visual)*, *Information Retrieval*, and *Query Language* concepts but not to the *Human Factors* concept (the triangle glyph to the South East of the central glyph) [125].

in the bottom row of Figure 3.13. While Gist Icons are an intriguing representation, users can only filter by one term at a time and more advanced combinations of query terms are hence not possible. In addition, the fact that only 50-100 documents can be shown within a frame means that users will have to traverse through many such frames while exploring a large collection of text documents. It is unclear if there is any further mechanism to make the traversal of Gist Icons more effective.

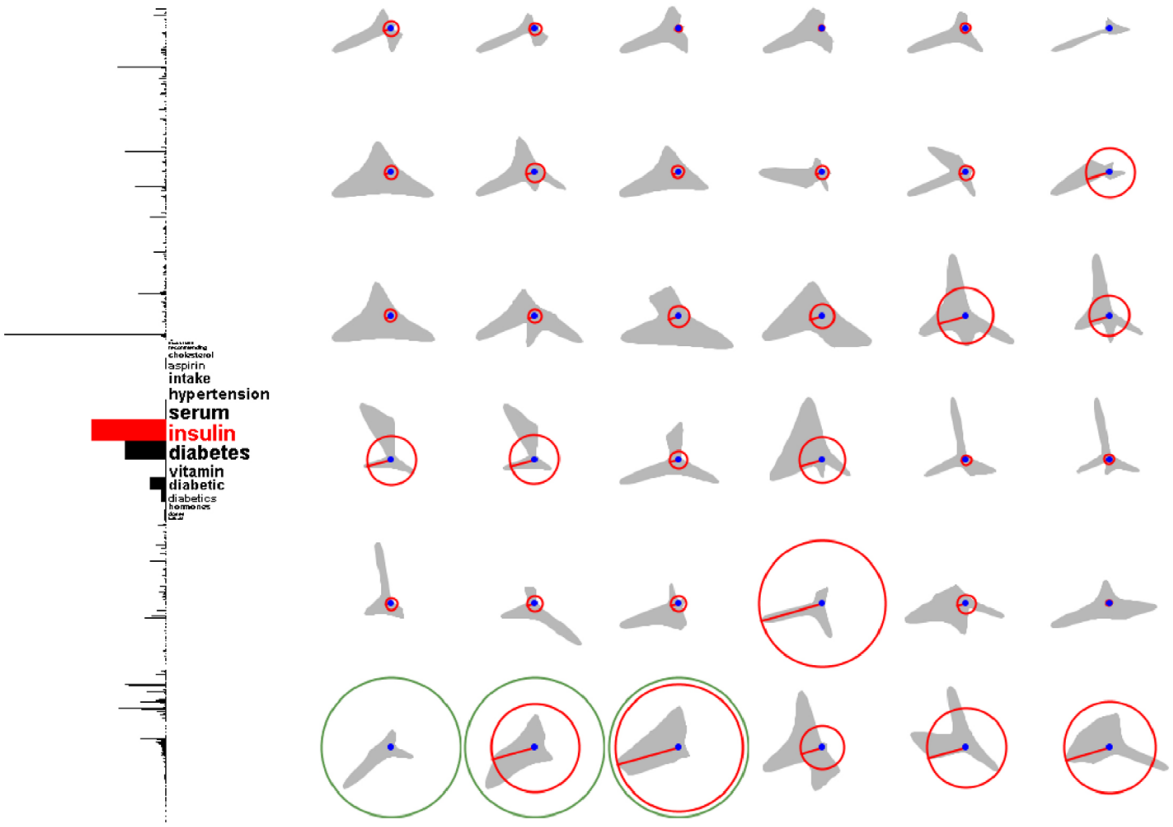


Figure 3.13: Gist Icons being used to represent patents [28]. Clicking on a term (e.g., "*insulin*") results in red circles and red lines indicating the relatedness of that term with respect to all documents containing it.

Unlike the filtering via query terms approaches highlighted thus far, the faceted filtering (or *faceted browsing*) technique supports filtering tasks via a set of flat or hierarchical pre-defined facets, which are usually used in combination with keyword search. The term "*faceted*" was adopted by the pioneering Flamenco project [159] to reflect the early idea of a colon classification system in 1933 from library science in which multiple classes are used to classify information items [102].

Figure 3.14 shows an example of a faceted browser being used to support filtering documents from a text collection. In this filtering paradigm, facets are categories characterizing items in a large collection [58]. Each facet has one or more facet values and each item may be associated with a subset of these values [21]. As such, this navigation paradigm usually requires rich metadata expressing relationships between facet values and resources. In this multi-step resource seeking process, users start with some initial constraint definition, inspect the initial results set, then continue with orienteering and refinement steps, and finish the process by closely examining the results set [57]. Within this process, users' selections of facet values result in either conjunctive or disjunctive queries executed on the resource collection, depending on the nature of the facets and/or the design of the applications. Query previews [100] are often used to indicate how many items are associated with each of the facet values (the numbers in parentheses next to the facets in Figure 3.14). As such, they help users avoid filtering by facet values that would return empty results sets [59]. The matching items, or focus items [21], are then displayed as results of the filtering process.

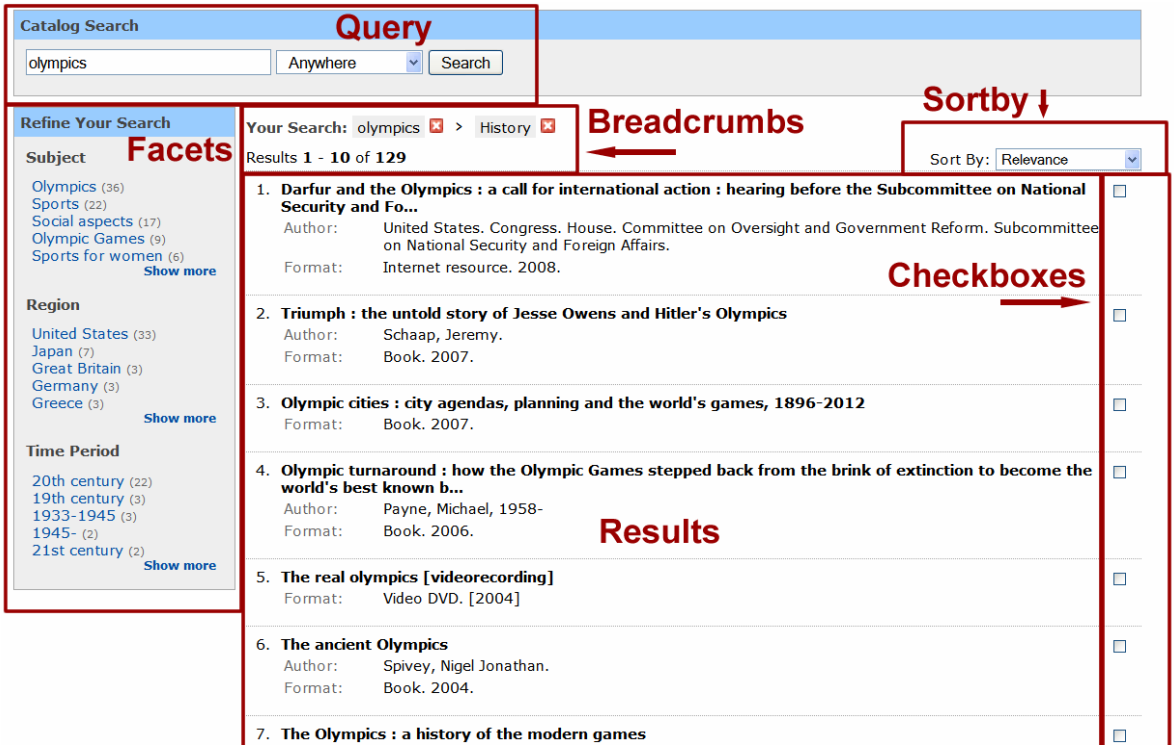


Figure 3.14: A example of a faceted browser on a text collection [80].

Considering the vast amount of research done on in faceted browsing, in this section we do not attempt to cover all advances. Reviews and thorough analysis of research work and commercial applications employing the faceted navigation paradigm can be found in [59] and [128]. Here we highlight a number of innovative faceted filtering systems which provide further support toward exploratory tasks by using more visual elements to show various basic statistics[2].

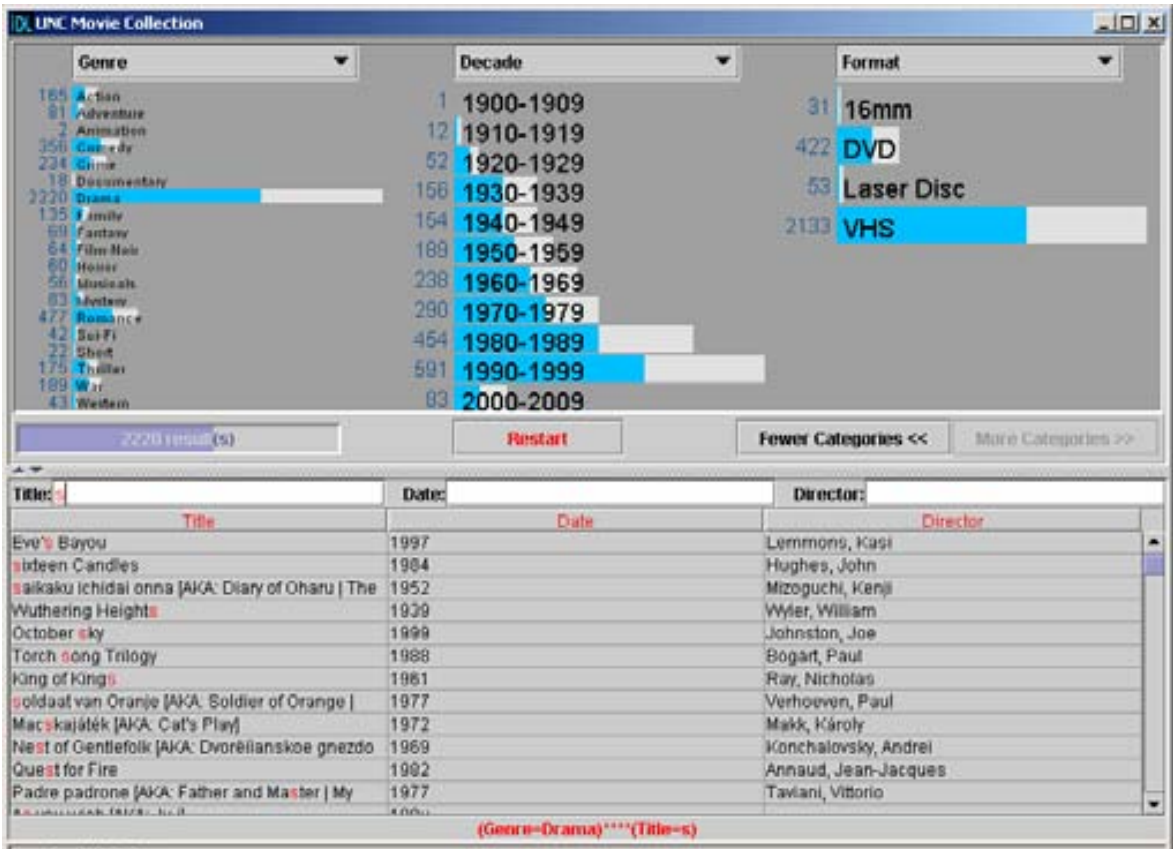


Figure 3.15: The RelationBrowser++ application [160].

For instance, the RelationBrowser++ application [160], as shown in Figure 3.15, visualizes the number of items matching the facet values using bar charts. Here the value "*Drama*" is selected in the facet "*Genre*"and the keyword "*s*" is used on the title field for searching.

[2]Even though the systems that we discuss here do not necessarily deal with text collections, they can generally be adapted to use with documents as units of analysis if the appropriate metadata are available. Furthermore, the free text components of information items in a collection tend to be used only for searching in combination with faceted filtering.

The bars with white backgrounds indicate how many items from a collection have been assigned to the categories, and the blue foregrounds imposed on those bars indicate how many items among them meet the criteria in each of the categories.

Having a similar goal as the RelationBrowser++ application, Elastic Lists [129] also show the proportions of items belonging to different categories. In Figure 3.16, the height of a list entry indicates the number of items having that facet value. In addition, the colors of list items are assigned such that brighter colors indicate that the proportion within the local context is much higher than that in the global context. For instance, in Figure 3.16, within the "*Gender*" facet the proportion of female who won Nobel "*peace*" Prizes is much higher than the proportion of female among all the Nobel Prize winners.

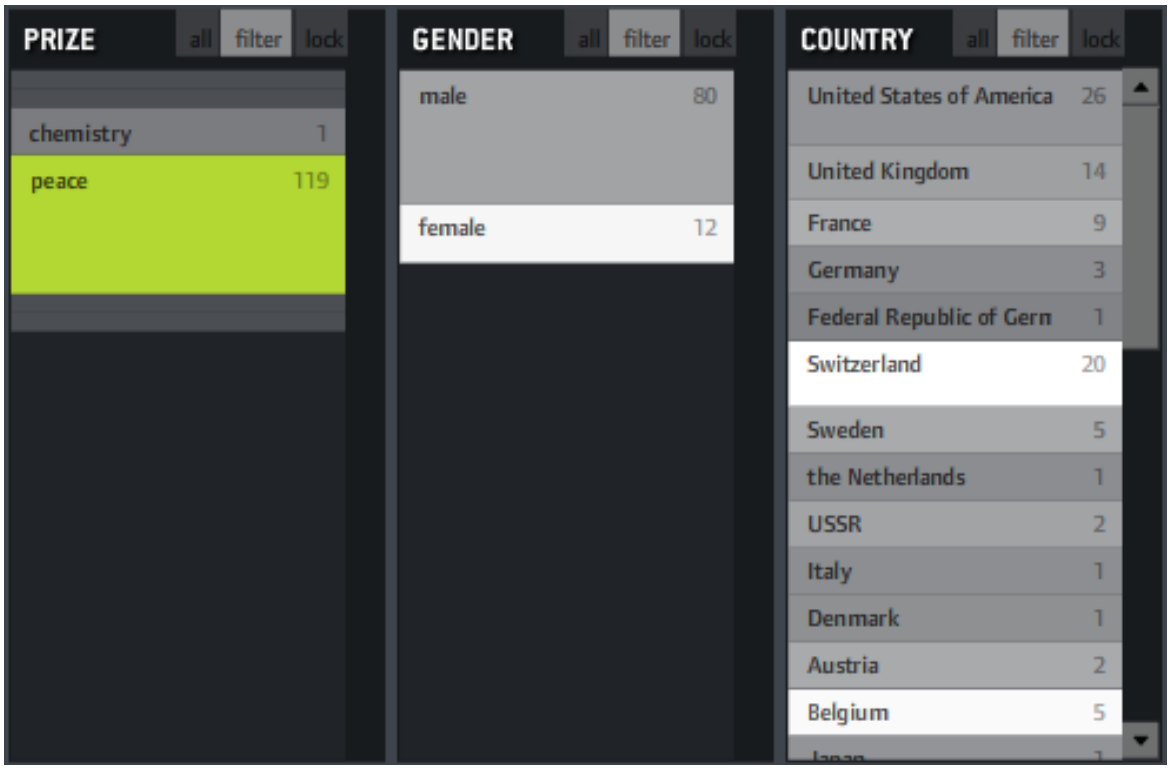


Figure 3.16: Elastic Lists for faceted browsing on the Nobel prize winners dataset [129].

It is also worth mentioning FacetLens [81], which focuses on visualizing the distribution of items within a collection over linear facets, such as time. Instead of using a stacked bar chart as in the RelationBrowser++ application discussed above, in FacetLens a more familiar horizontal timeline is used. FacetLens, however, uses circles to represent focus items, and

group them into rectangle panels. This layout, however, is not space efficient (see [81] for screenshots).

Faceted browsing research also receives a lot of attention from the Semantic Web community, with works such as mSpace [114], /facet [62], BrowseRDF [96] and Aduna AutoFocus [42] supporting filtering of semi-structured data.

In mSpace [114], a column-based interface is used, in a similar way to the Elastic Lists interface but without the proportion information in visual form, and users are allowed to filter via only one facet value at a time. /facet [62], on the other hand, is an initial effort which focuses on providing support in narrowing down which facets to use in case of them being encoded in a large tree structure. However, the authors of /facet themselves stated that there remains many challenges in classifying facets into hierarchy to deal with items associated with many facets, in identifying properties and classes that are most beneficial to users [62].

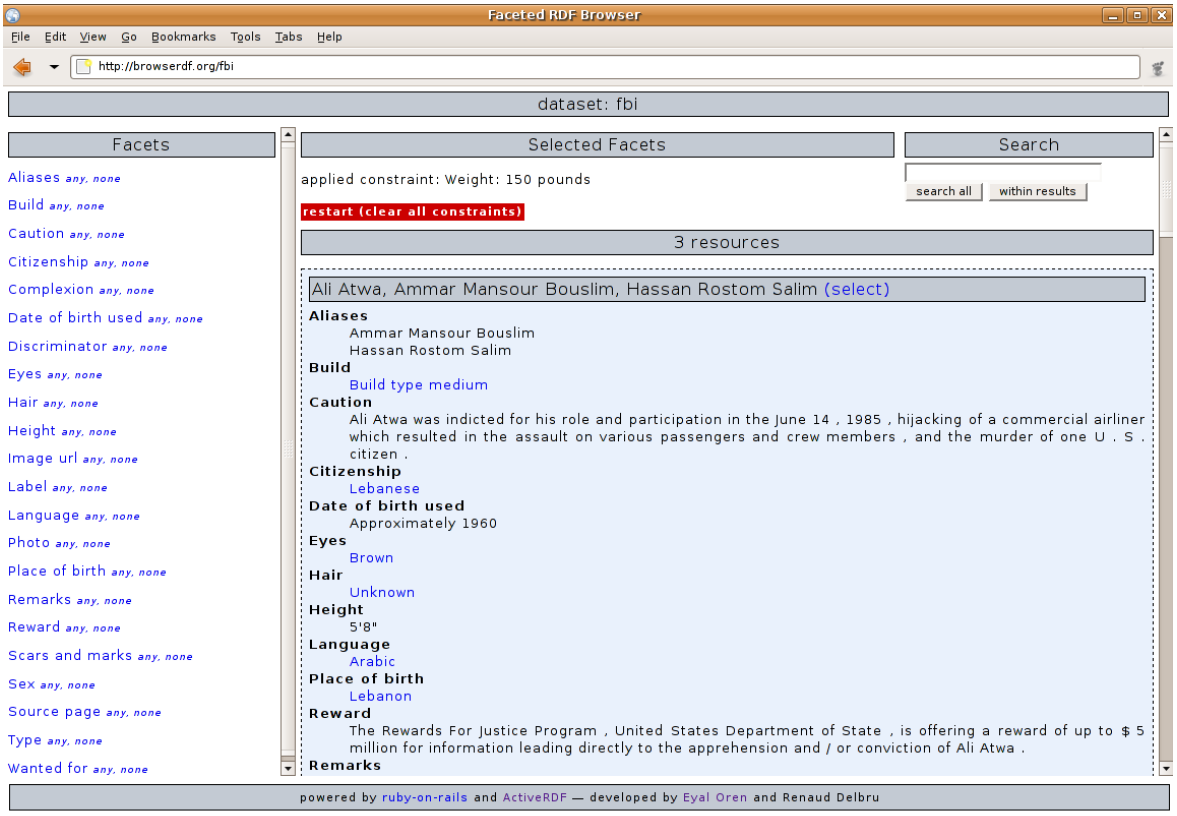


Figure 3.17: The BrowseRDF application [96].

The issues identified in the /facet work are initially addressed within the BrowseRDF

application [96], as shown in Figure 3.17. An automatic facet ranking approach, which is based on the decision tree data classification paradigm, is employed in BrowseRDF to select facets that enable efficient navigation through an RDF dataset. A number of metrics, such as predicate balance, object cardinality, and predicate frequency, are used in combination to rank facets (c.f. [96] for further details on how these scores are formulated). Even though this approach is well grounded, its authors also note that there is room for further research as the ranking does not always correspond to the intuitive importance people assign to some facets [96].

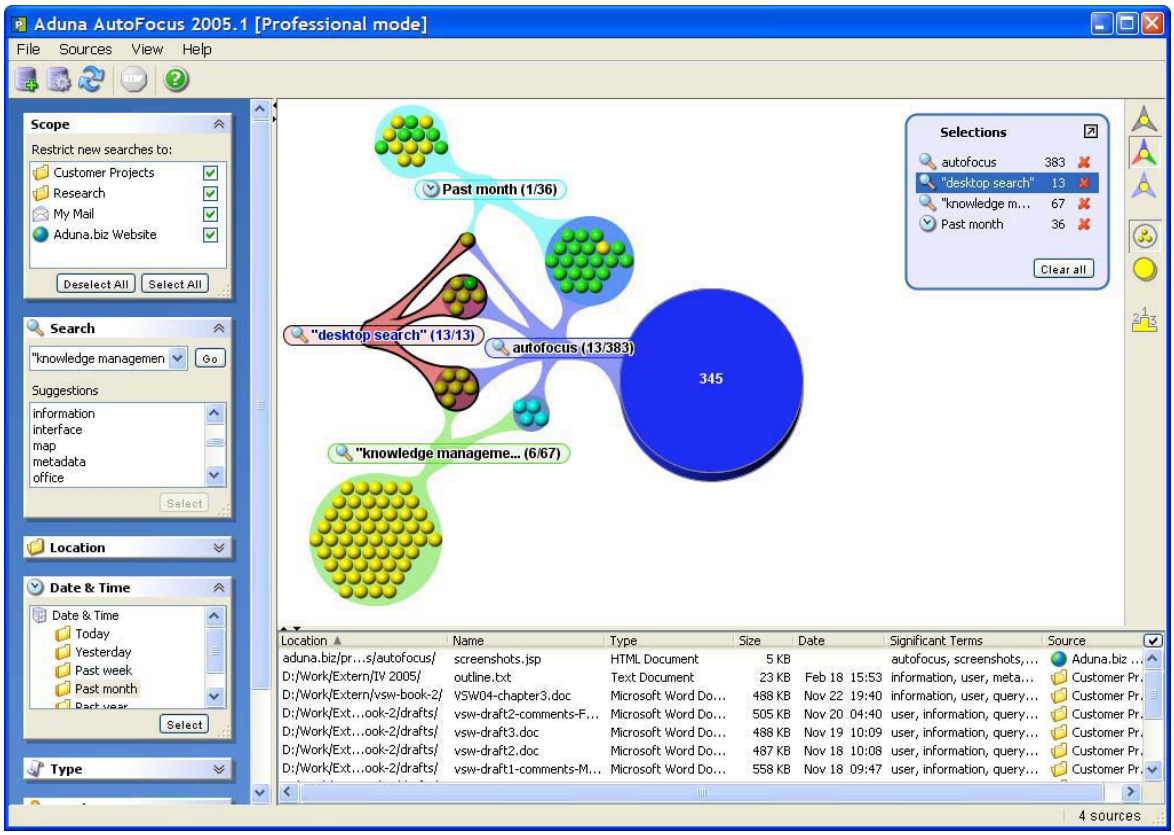


Figure 3.18: Aduna AutoFocus [42].

Apart from the above research efforts, Aduna AutoFocus [42] is a commercial desktop search application that takes advantage of previous works in visualization such as the InfoCrystal [125] and VIBE [95] to support faceted browsing of documents on a user's desktop. In AutoFocus, as shown in Figure 3.18, users can define the scope of the search space (the "*Scope*" panel in the upper left hand side corner of Figure 3.18), and then specify either

facet values or query terms to filter for relevant documents. In the resulting display, colors encode the type of documents (e.g., files are yellow spheres, web pages are green spheres, etc.). Moreover, documents matching different combinations of the selected facet values or query terms are grouped together, and displayed in such a way that resembles a combination of the VIBE [95] and InfoCrystal [125] approaches discussed above. For instance, in Figure 3.18, there are four labels representing the four filtering criteria, and if a group of spheres (documents) is linked to a single label, they represent documents matching that criteria only (e.g., the group of yellow spheres on the lower left corner of Figure 3.18 are for documents matching the keyword "*knowledge management*"). Meanwhile, if a group of spheres is visually connected to more than one label, it corresponds to documents matching all the conditions represented by those labels. For instance, in Figure 3.18, the group of four blue spheres to the North East of the previous group represents documents matching both "*knowledge management*" and "*autofocus*". It is unclear, however, how this graph-based layout approach can cope with a relatively large number of facet values or query terms, as there will be many edges crossing each other to represent documents matching different combinations of the facet values and query terms being used for filtering. In addition, the author of AutoFocus also emphasizes the need to enable users to define their own meaningful categories or taxonomies to organize their information space based on their own interests [42].

Another major development in the faceted browsing area is the lightweight structured data publishing Exhibit framework[3] [64]. While not a contribution to the faceted browsing paradigm research *per se*, the Exhibit framework essentially enables casual users to publish their (typically small) structured dataset on the Web via a faceted browsing interface at ease without the need for any server setup and configurations that would require considerable technical knowledge and effort.

As popular as it is, faceted browsing is not without its problems. Notably, in [59], Hearst notes that: "*There are some deficiencies with the faceted paradigm. If the facets do not reflect a user's mental model of the space, or if items are not assigned facet labels appropriately, the interface will suffer some of the same problems as directory structures.*" The concern

[3] `http://www.simile-widgets.org/exhibit/`

of a potential mismatch between a faceted browsing system and a user's mental model is particularly relevant to our work. We will discuss our proposed approach in this regard in Chapters 4 and 5.

3.4 Visual Analysis of Document Structures and Term Distributions

All of the approaches described in the previous section treat a document as a unit of analysis, i.e. a single feature vector is used to characterize the whole text of a document. While this treatment allows for analysis of inter-document similarities, a lot of information about the internal distribution of entities across a document is disregarded [74]. In this section, we highlight innovative approaches used in the visual analysis of document structure and term distribution, which can provide users with useful details on demand while analyzing a document.

TileBars [56], as shown in Figure 3.19, is one of the early and best-known visualizations to show the distribution of query terms within documents. TileBars is a visual metaphor that enables users to decide which documents and which parts of a document to focus on based on the distribution of query terms within documents retrieved from a text search. In this metaphor, a document is preprocessed and segmented into sequential logical units by the TextTiling algorithm [55] and these units are visually represented by small square blocks, colored using a variation of grayscale depths to indicate frequencies. A document is then shown as a rectangle encompassing these blocks of logical units [56]. As such, a TileBars visualization enables users to see many features at once despite its compact visual representation: (1) relative lengths of retrieved documents, (2) locations and frequencies of query terms within a document and (3) whether occurrences of query terms coincide with logical units of the text. It is interesting to note that even though research in the perceptual psychological literature suggests that grayscale variations lead to more accurate information, further research shows that "*the aesthetic preference for color outweighs the need for accuracy*" [59]. A TileBars display, however, becomes increasingly difficult to read when there

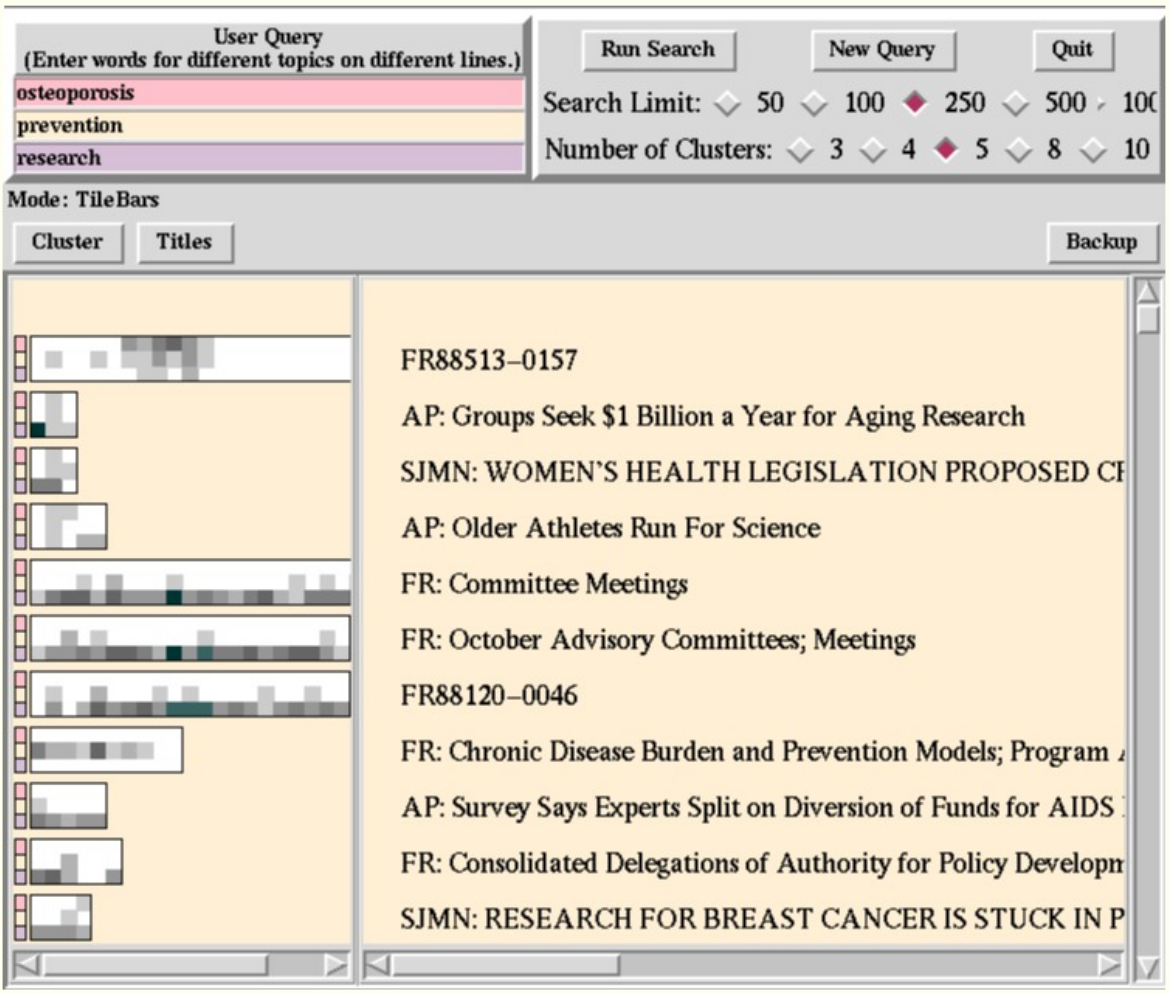


Figure 3.19: A TileBars visualization [56].

are more rows in it. As a result, this would hinder quick interpretation of the distribution of terms within a document.

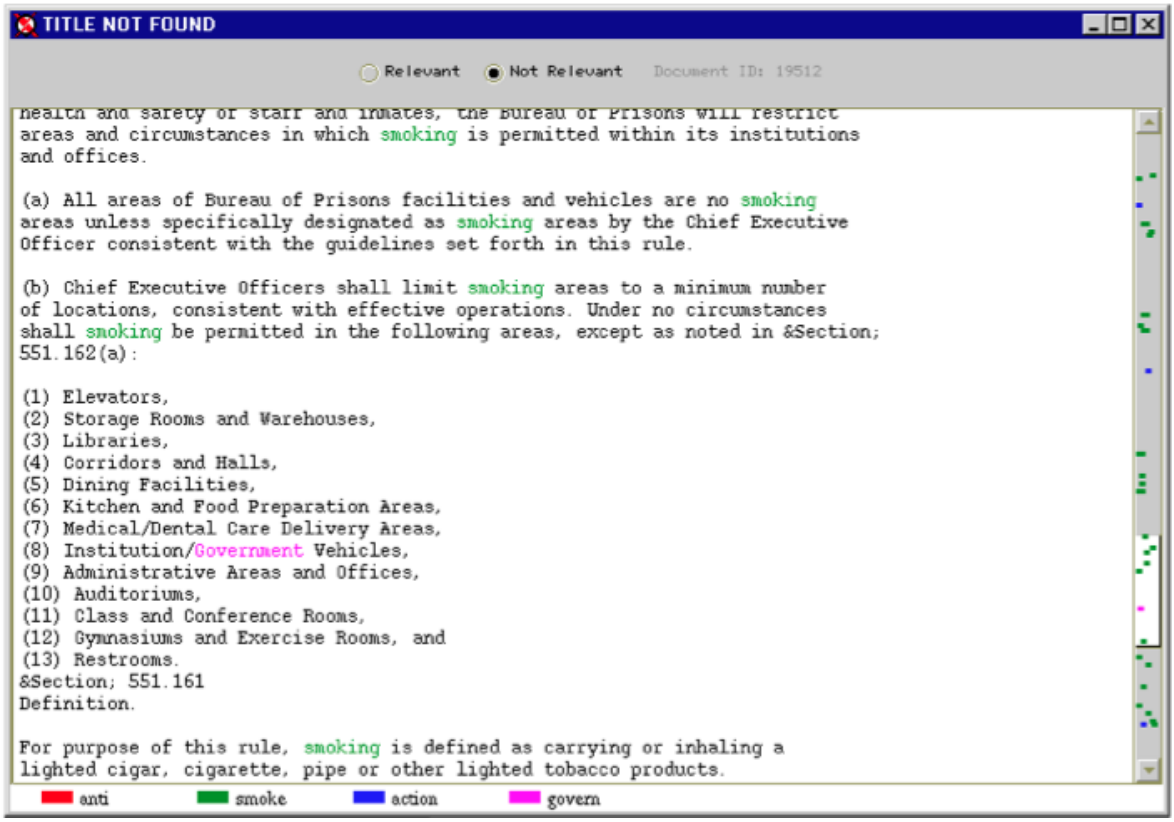


Figure 3.20: A scrollbar-based term distribution visualization [11].

In [11], Byrd proposes a visual approach that focuses on a navigation support within a text document. This is a visualization that makes innovative use of the scrollbar area within a document viewer. In this visualization, as illustrated in Figure 3.20, each query term is assigned a color in the legend at the bottom of the document viewer, and each occurrence of a query term is highlighted in the scrollbar area. The scrollbar handle has a white background and corresponds to the text portion being displayed in the viewer. As a result, the scrollbar area is a miniature view of query term distribution within a document [11]. In comparison with TileBars, this visual representation does not focus on comparison of term distribution across documents. It also has the advantage that both the term distribution and the textual contents are immediately available in the same view, hence does not require any further clicking by users to jump to the text as in TileBars. However, when a chunk of text contains

many of the query terms, the scrollbar might be overcrowded with many dots representing mentions of such terms.

A more recent work on term distribution visualization is proposed in [115] which is inspired by TileBars and employs histograms at different levels of detail in a Focus+Context way. As shown in Figure 3.21, there is a color-coded histogram for each term of interest within a document. The areas under the histograms can either be filled or unfilled with matching colors. In the filled version (Figure 3.21) a solid line is added on top of the curve to allow one to see details that might be obscured by color blending [115]. In addition, the visualization allows users to read it at different levels of detail by brushing an area of interest within the histogram area, as illustrated in Figure 3.21. While this focus+context histogram visualization is visually-appealing, it is not clear whether color blending might cause any issues with understanding the visualization in detail. The authors themselves state that color blending and weaving would be the focus for future improvement [115].

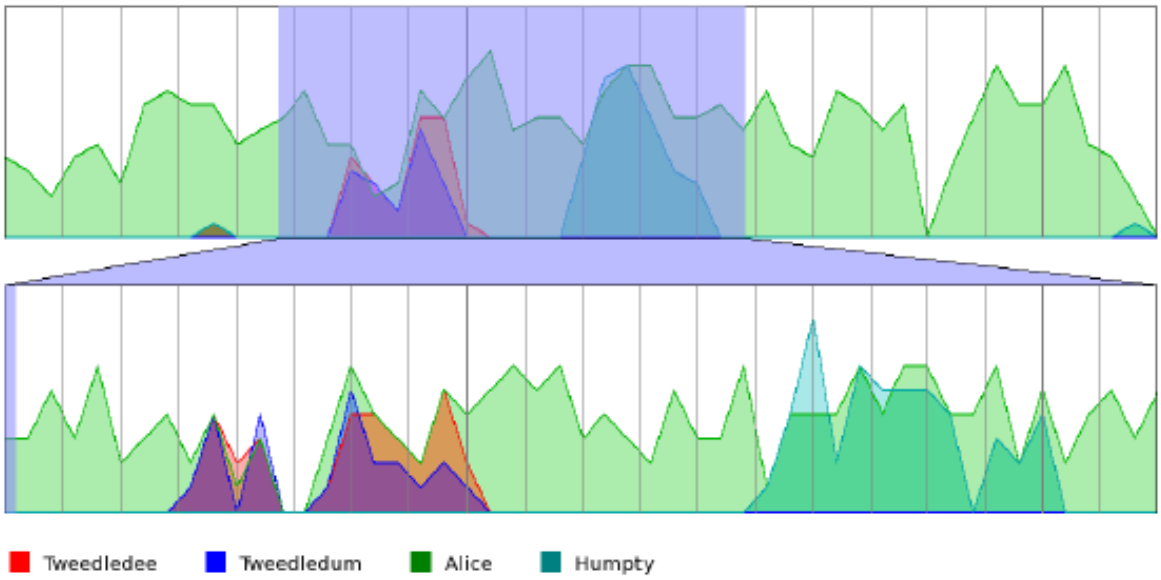


Figure 3.21: A term distribution visualization with Focus+Context views [115].

Seesoft [35, 37] is a visualization technique that is useful for graphically depicting software source code and word usages in a text corpus. Seesoft shows a text file as a vertical column and each line as a color-coded row within the column [35]. Full text view of a small portion of text can be shown on demand. The spatial pattern of color helps to illustrate the

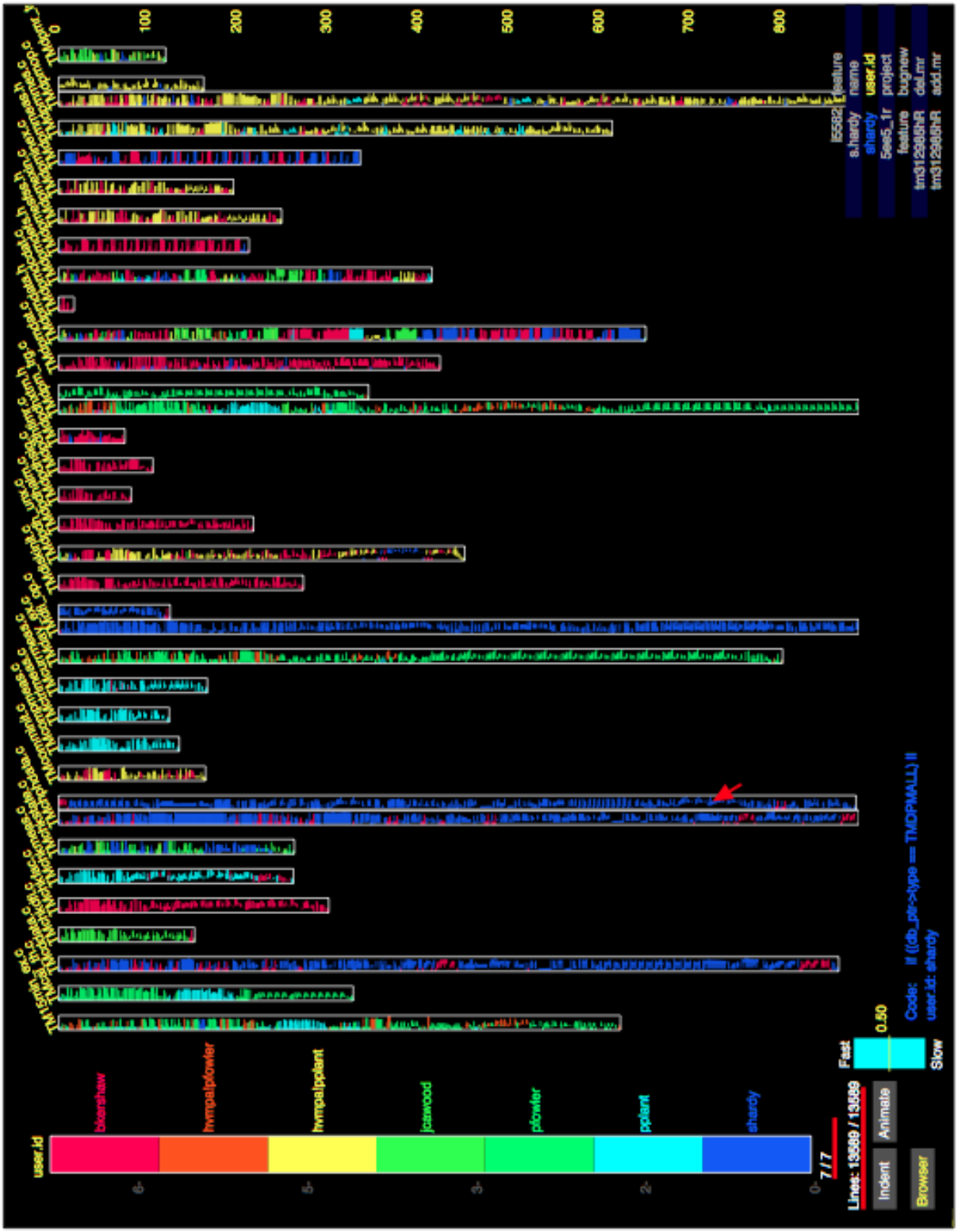


Figure 3.22: A visualization of code written by seven different programmers within source files with Seesoft [35].

distribution of statistics of interest (here statistics can be categorical such as authors, continuous such as documents' time stamps, or binary such as whether a line has been changed) within a text file. As shown in Figure 3.22, the Seesoft visualization is a compact and aesthetically pleasing representation that can help to abstract textual details away [59]. A known limitation of Seesoft is that as an entire line is color coded with an attribute value (e.g., the person who made changes to a line of code), it is not easy to visually depict mentions of entities at a finer grain level. For instance, if used on novels, Seesoft cannot highlight the mentions of two different characters appearing on the same line of text with this layout [59].

Related to the Seesoft visualization is Compus [38]. Instead of focusing on free text, Compus is developed to deal with corpora of documents that are encoded using the XML/TEI format to describe paratextual phenomena such as abbreviations, insertions, deletions, corrections or unreadable portions of text [38]. Figure 3.23 shows a Compus visualization of 100 XML documents. Each XML element is assigned a color and is visualized based on the character offset where the element starts and ends within a document [38]. As shown on the right side of Figure 3.23, each document is visualized as a 5-pixel wide rectangle containing wrapped lines of text. Since XML elements can be nested, the inner most elements are visualized over the text span that they contain. While Compus can enable users to compare the composition of documents based on a set of standard markups, it is unclear how Compus can scale up to larger collections of text. In addition, there was no usability evaluation conducted on this visualization to see how well it can support users in analyzing such structured documents.

Visual Fingerprint [74] is a visualization that enables users to gain an understanding about various characteristics of a text document at different levels such as words, sentences, paragraphs, chapters, etc. In a visual fingerprint, each unit of analysis (a block of text) is represented by a colored square, sequentially lined up from left to right, top to bottom, with the filled colors reflecting values of statistical measures, e.g., sentence length, vocabulary richness, etc. [74]. Visual fingerprints are useful in literature analysis, for instance, to determine authorship attribution (see [74] for examples), and in other more general purposes such as detecting lengthy sentences in a publication so that readability can be improved, as shown in

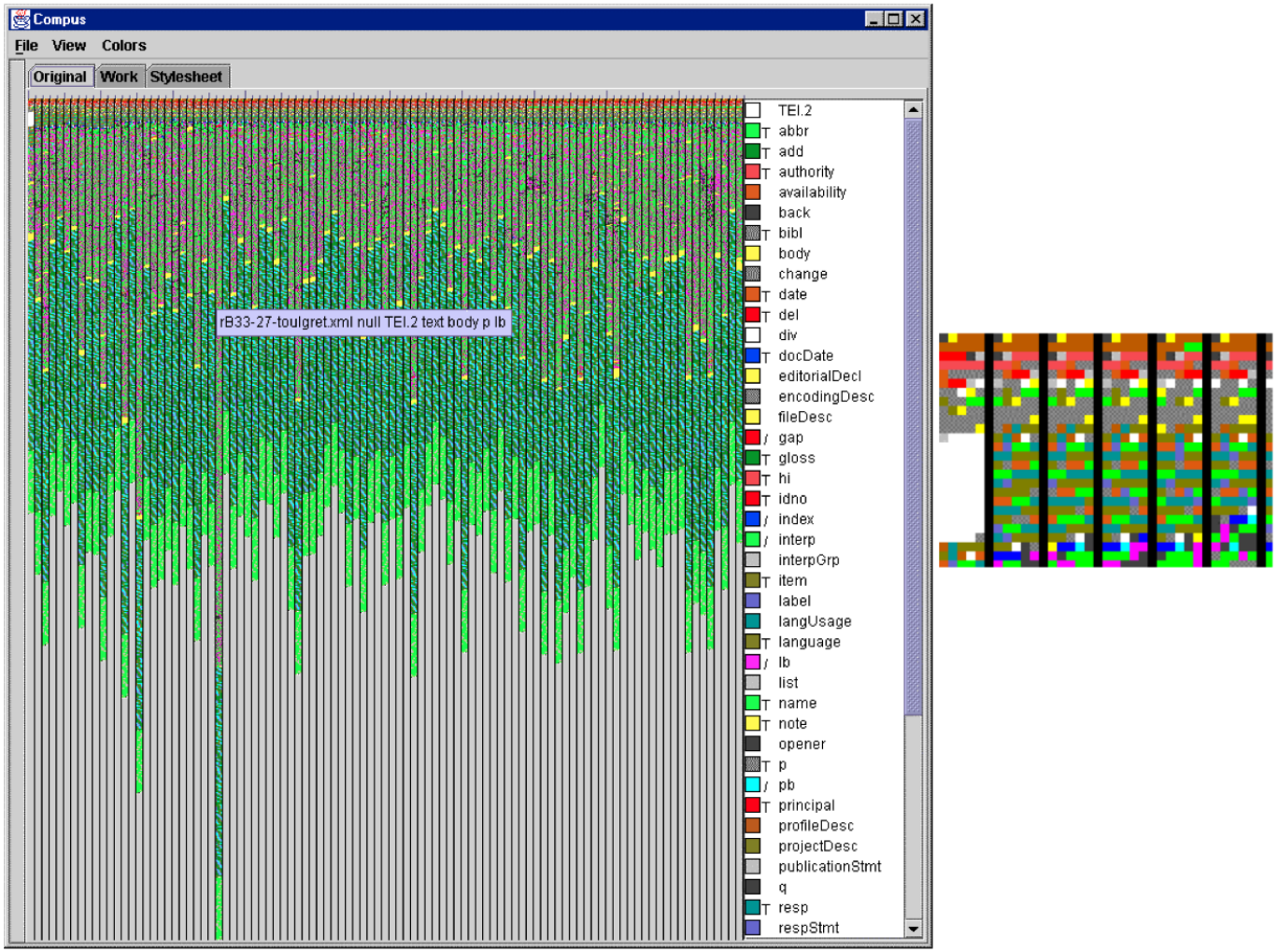


Figure 3.23: A Compus visualization: a collection view on the left and a zoomed in view on the right [38].

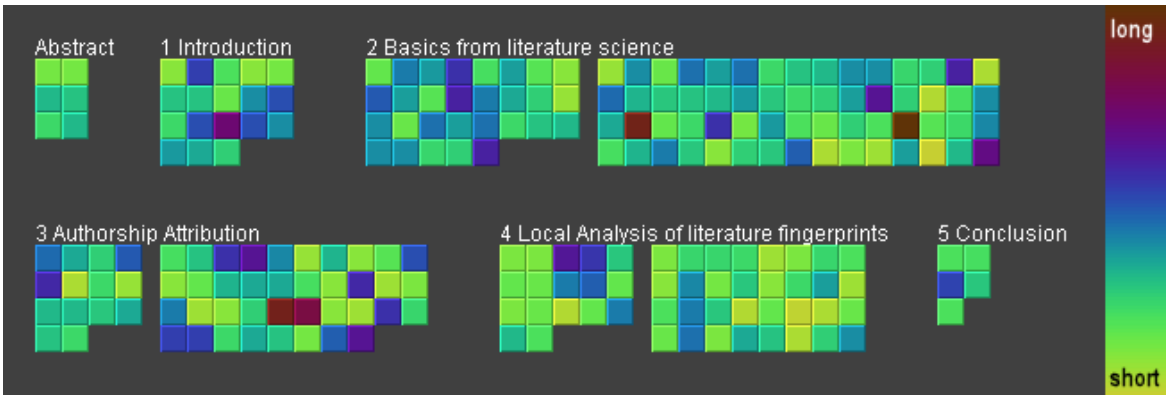


Figure 3.24: A visual fingerprint showing the lengths of sentences within a document [74].

Figure 3.24. These text fingerprints are also used in visual evaluation of document summarization and opinion analysis, in which they visualize the interim results generated from the analytic process and hence enable analysts to make changes to obtain improved results [94].

Another line of research employs machine learning methods to derive labeled structure of documents. One of the main challenges for such automated document structure analysis tasks is the heterogeneity of document types, which might hinder the application of rule-based approaches developed for a specific category of documents toward various other types of documents [131]. The results of document structure analysis can be used in many ways, for instance, to give users a quick understanding of the distribution of subject matters within documents, or to augment queries with additional information about the logical structure of documents to become more expressive queries (for example, using the query "*introduction: engineering*" to search for documents containing the term "*engineering*" in the introduction section [131]).

In [131], a document structure analysis method is proposed which combines a machine learning technique with an interactive visualization. In this method, a standard data classification method assigns pre-defined labels to lines of text based on a set of layout and formatting features [131]. Users are involved in the training process in that they can directly verify and correct misclassified labels. An example outcome is shown in Figure 3.25, whereby lines of text within a document are color-coded to indicate the category they belong to and users can manually change incorrect assignments by selecting the right label in

Figure 3.25: Automated visual analysis of document structure [131].

a context menu. This approach is a good example of a visual analytics system in which user interaction can help improve the analytical outcomes.

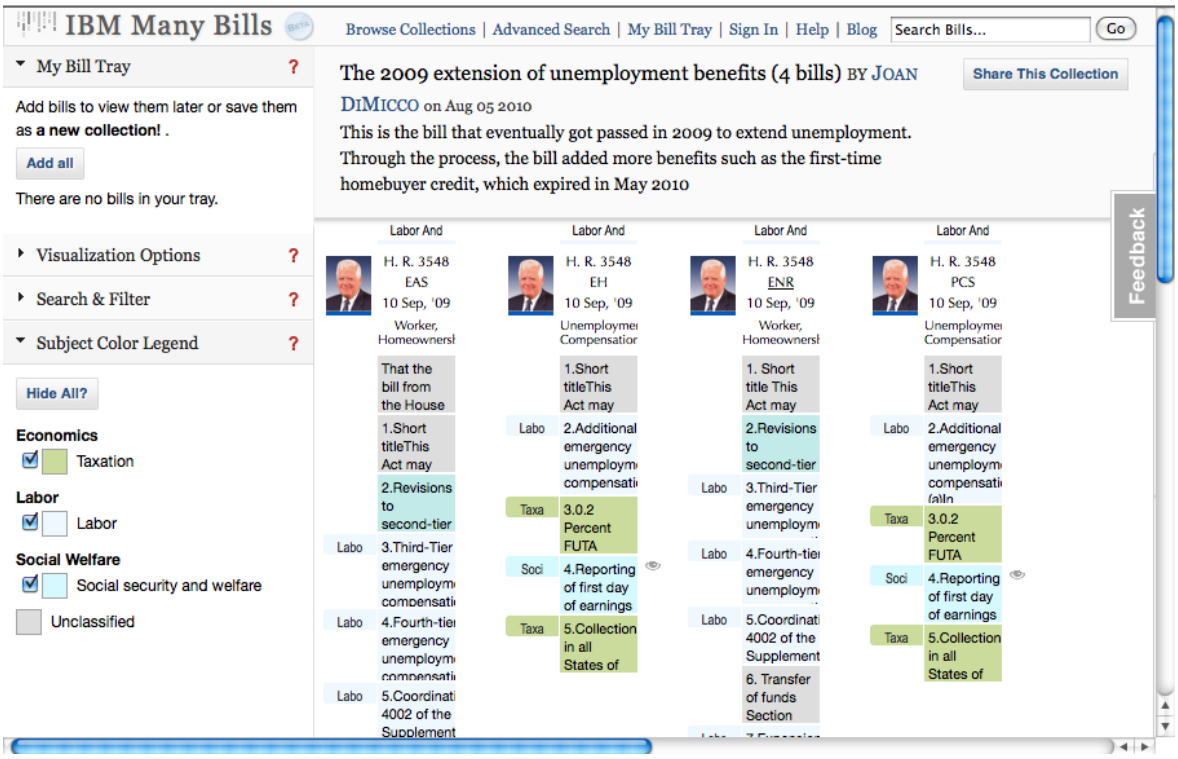


Figure 3.26: A structure visualization of four example bills in the Many Bills application [3].

The latest addition to the document structure visualization area is the Many Bills[4] application [3]. This web application is designed specifically to address the challenges of length, complex language and obscure topics while navigating US legislation [3]. One of the main goals of Many Bills is to provide an interface that shows different levels of content abstraction to assist users in understanding various aspects of the text. Many Bills employs a text classification algorithm to label individual sections within bills with subjects from a predefined set. These labels are then used to visualize the topical substructure of bills using color-coded, vertically stacked blocks, as shown in Figure 3.26. In addition, blocks with assigned subjects that are semantically related are color-coded similarly using only one color. This visualization is helpful in providing a guidance to users in terms of deciding whether a bill or some parts of it are interesting to read [3]. Furthermore, Many Bills has the advantage

[4] `http://manybills.researchlabs.ibm.com/` - Last accessed date: 18 August 2011

of being intuitive due to its simple visual representation. However, the assignment of only one subject label per section might be restrictive, as a section might actually be about more than one topic.

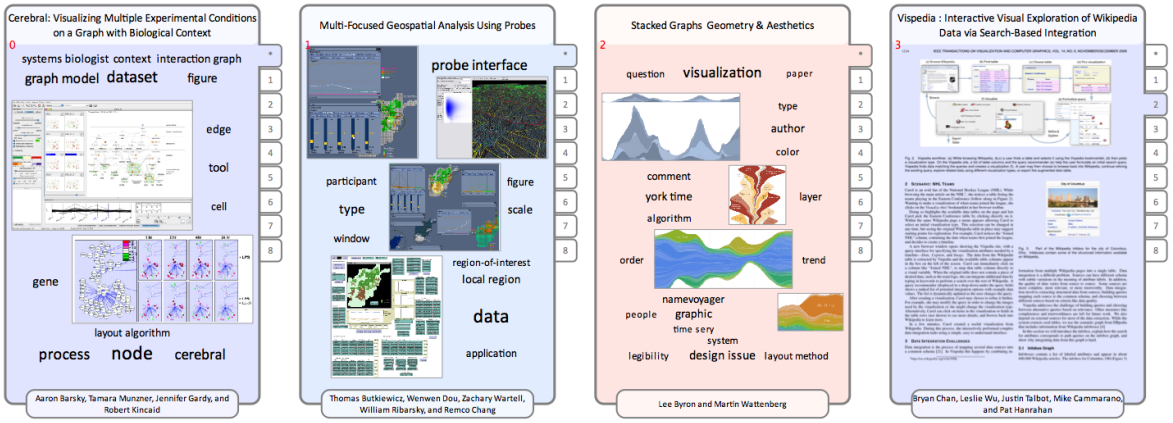


Figure 3.27: A set of Document Cards showing key terms and images from four scientific publications [132].

It is also worth mentioning Document Card [132], which is an innovative visualization technique to show a compact summary of a document in a fixed size, thumbnail-like representation. Given a document, the proposed approach extracts key terms and important images from it. Images that contain key terms in its caption and referencing text are considered important. For more detail on how the key phrase and image extraction is carried out, as well as, the layout approach, interested readers are referred to [132]. Figure 3.27 shows a number of Document Cards representing summaries of four scientific publications. The tabs on the right hand side of each document enable users to select to view Document Cards in either page mode (the fourth one in Figure 3.27) or document mode (the first three in Figure 3.27). The extracted title and authors of each document are shown at the top and bottom of a Document Card respectively. The key advantage of Document Cards is that they present key information in a compact representation, which makes it possible for them to be displayed on different devices, such as mobile devices with limited screen estate, or large displays that can show multiple Document Cards at once. However, the usefulness of this visual summary metaphor might be limited to documents containing a number of figures. In the case of text-only documents, Document Cards will become tag cloud-like displays of key

terms extracted from documents.

Overall, as text is usually written from left to right, from top to bottom in many languages, the approaches described here share a commonality in that they take advantage of such a linear structure of text within a document to show relevant information. A visualization of document structures and term distributions can enable users to quickly understand which parts of a document are about what, hence can tailor their focus accordingly.

3.5 Visual Concordance Analysis

Visual concordance analysis can serve to support users in having a quick understanding of how terms are used in text. These visual perspectives on a text are potentially helpful in many ways as they can act as visual summaries and subsequently provide jumping-off point for close reading [142]. The usefulness of concordance analysis has led to various research approaches and commercial software.

The Concordance software[5] provides a tabular view, with the seed term in the middle column, and the text written before and after it shown in the side columns. While this simple display can be used to support detailed analysis of concordance within a single document, it can be taxing to traverse a long list of concordances when the term in question appears too many times. Furthermore, this layout does not allow for an aggregation of concordances so that users can investigate at document level.

A more visually oriented tool called TextArc[6] [97] shows the frequencies and distributions of words within a text via a spiral layout. Given a document, the entire text is drawn line by line in an eclipse and in a clockwise direction starting from the top center. Anchors such as headings, chapter breaks, etc. are also drawn with an aim to retain the typographic structure of the text. The exact same text is drawn again in an inner eclipse, but word by word. Words that appear more than once are drawn only once at the center of all the positions around the eclipse where they should be placed [97]. This interesting placement scheme enables users to tell whether a word is evenly distributed in a text (it is so if it is more toward the center) or

[5] `http://www.concordancesoftware.co.uk/` (Last accessed Jan 12, 2011)
[6] `http://www.textarc.org/` (Last accessed Jan 12, 2011)

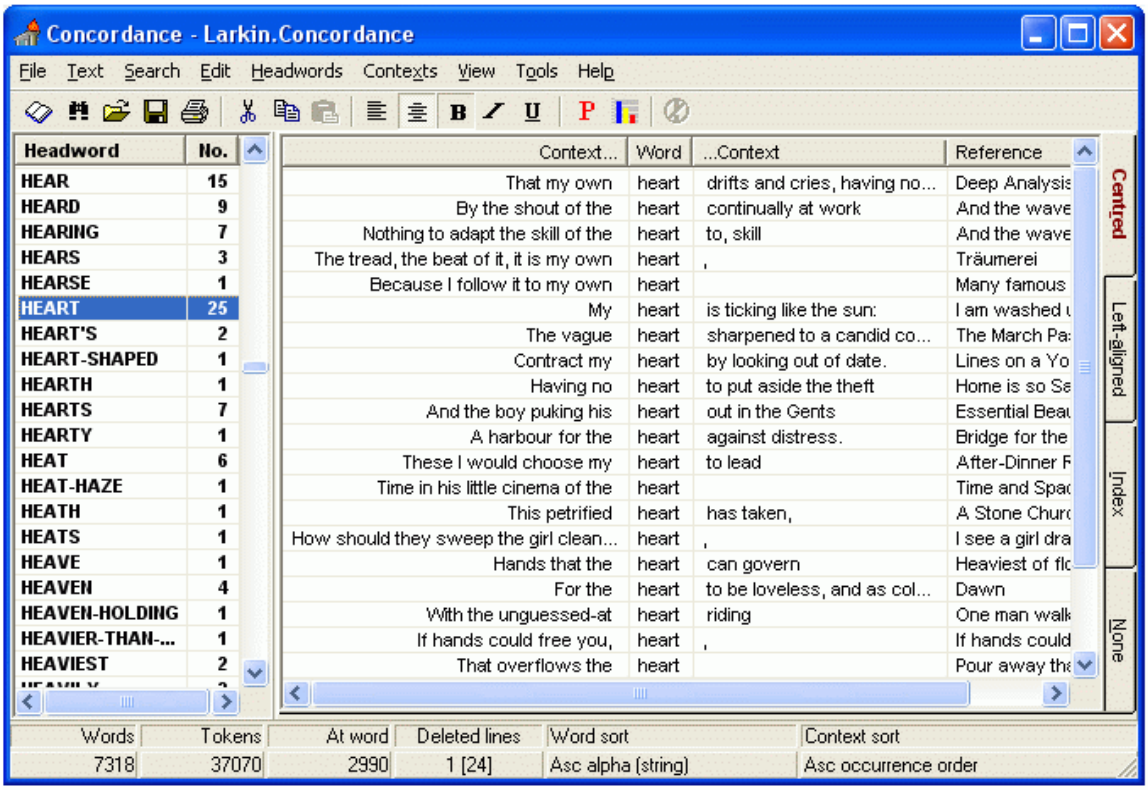


Figure 3.28: A Concordance visualization with "*heart*" being the seed term.

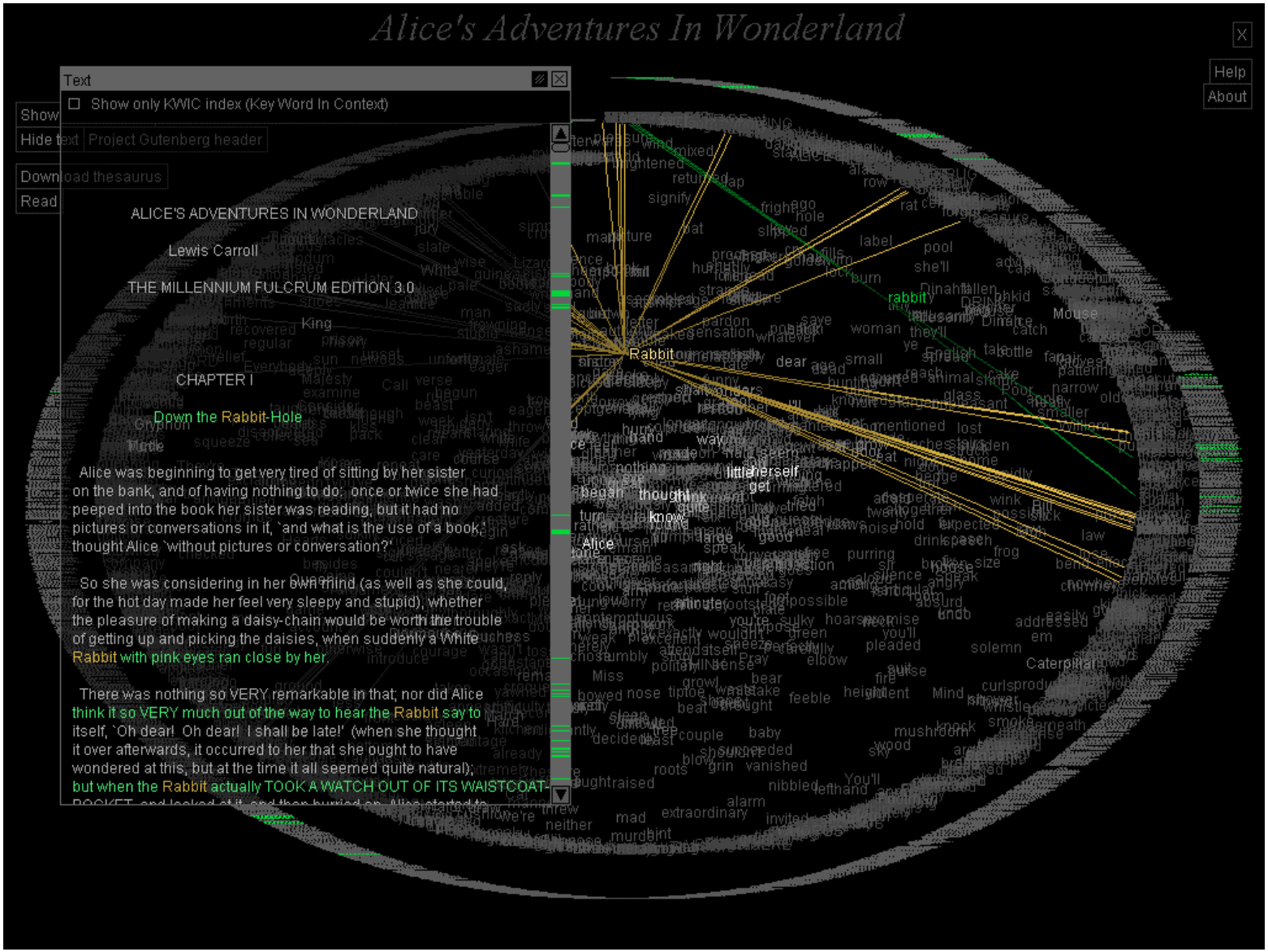


Figure 3.29: A TextArc visualization [97].

it is pulled toward a particular section. Figure 3.29 shows words in Lewis Carroll's *Alice's Adventures in Wonderland*. The word "*rabbit*" is highlighted and links to lines containing it are made visible. While being an interesting visualization tool, in TextArc no text analysis is performed to abstract away the details. As a result, TextArc is cluttered with too much text which affects legibility. In addition, it is not easy in TextArc to see the contexts in which a term appears as a full text view is necessary to get that information.

Digital humanities researchers also use concordance analysis in their effort to understand the changing semantics of terms over time[7]. Other works in visual concordance analysis focus on visualizing patterns in different ways. Of importance among them are Word Tree [151], Phrase Net [142], and FeatureLens [32].

Word Tree [151] is a simple yet visually engaging representation that shows term usages in context, with a focus on repetitive patterns. This representation is a visual and interactive version of the "*keyword-in-context*" technique [151]. Given a seed term, a tree structure is used to show which words are most frequently used right after it within the text. Figure 3.30 shows a word tree for the seed term "*if*" in *Romeo and Juliet*. In part A, we can see that "*if*" is most frequently followed by "*thou*". By zooming in to a particular branch of the tree, users can further study other terms within the usage contexts of the original seed term as shown in part B, as well as switching the focus of analysis entirely to a different seed term as shown in part C of Figure 3.30. An interesting addition to the current design of word tree would be a version that enables visualization of the wordings used both before and after a seed term as enabled in the Concordance software mentioned above.

Phrase Net [142] is a visualization technique to display relationships between words matching a pattern in a document. These relationships are shown using a network whose nodes are words matching the pattern and directed edges show the links between them, weighted by the number of times the pattern was matched. As the resulting graph can be too big to be easily comprehensible by users, edge compression using topological equivalence is carried out to group nodes that have identical neighbors [142]. In this visualization, terms with high out-degree to in-degree ratios are encoded with darker colors to enable users

[7] `http://www.cforster.com/2010/01/mining-obscenity-iii-failure-with-visualizations/` (Last accessed Jan 12, 2011. Note though that although the work suffers from noisy/missing data, it illustrates well the attempt to analyze usage contexts to understand the evolving semantics of a term.)

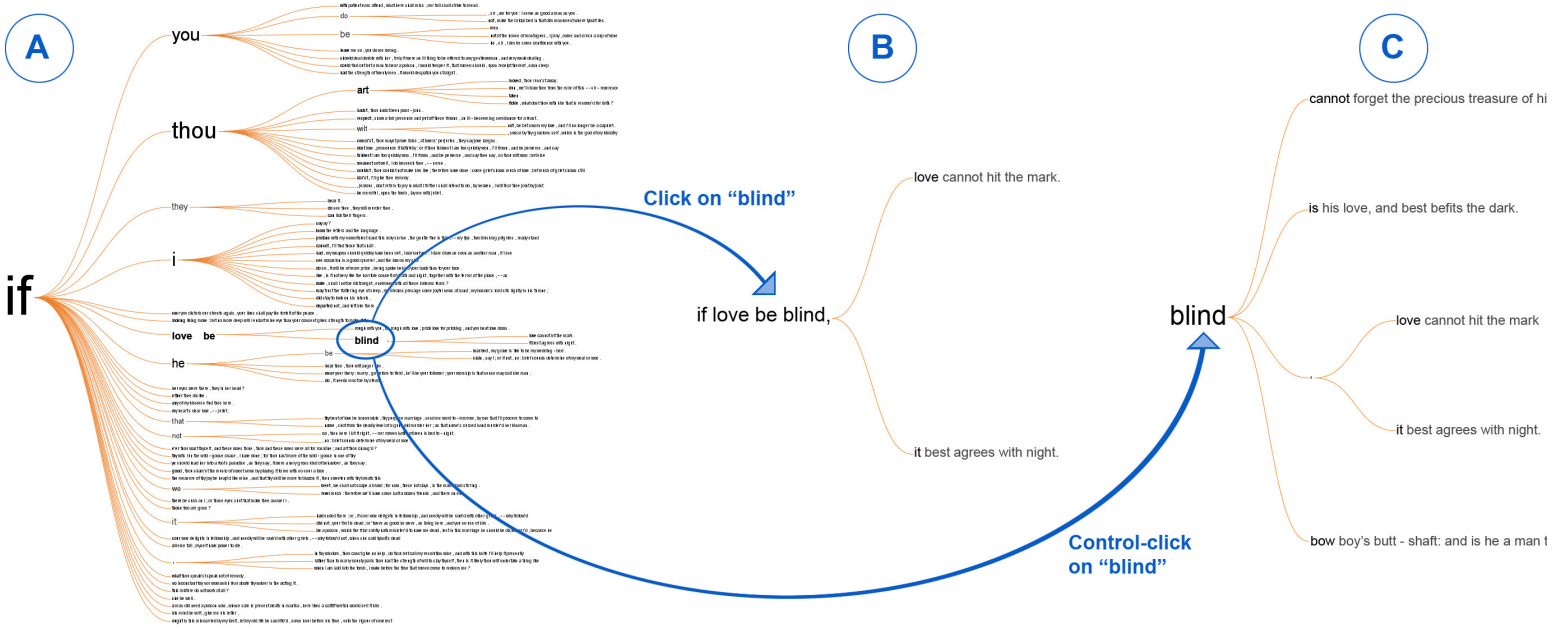


Figure 3.30: A Word Tree for the seed term “*if*” in *Romeo and Juliet* [151].

Figure 3.31: Phrase Nets for Jane Austen's *Pride and Prejudice*, with the patterns "*X and Y*" and "*X at Y*" being used on the left and the right hand side respectively [142].

to quickly identify terms that only occur in the first or last part of the matching pattern. Figure 3.31 shows two examples of Phrase Net when applied to Jane Austen's *Pride and Prejudice*, with the patterns "*X and Y*" and "*X at Y*" respectively. The left Phrase Net mostly shows relationships between main characters and the right Phrase Net displays relationships between locations. Similar to Word Tree, Phrase Nets do not link back to the places within a text that the seed term appears. Hence, a term distribution view is missing even though it would be helpful for analysis.

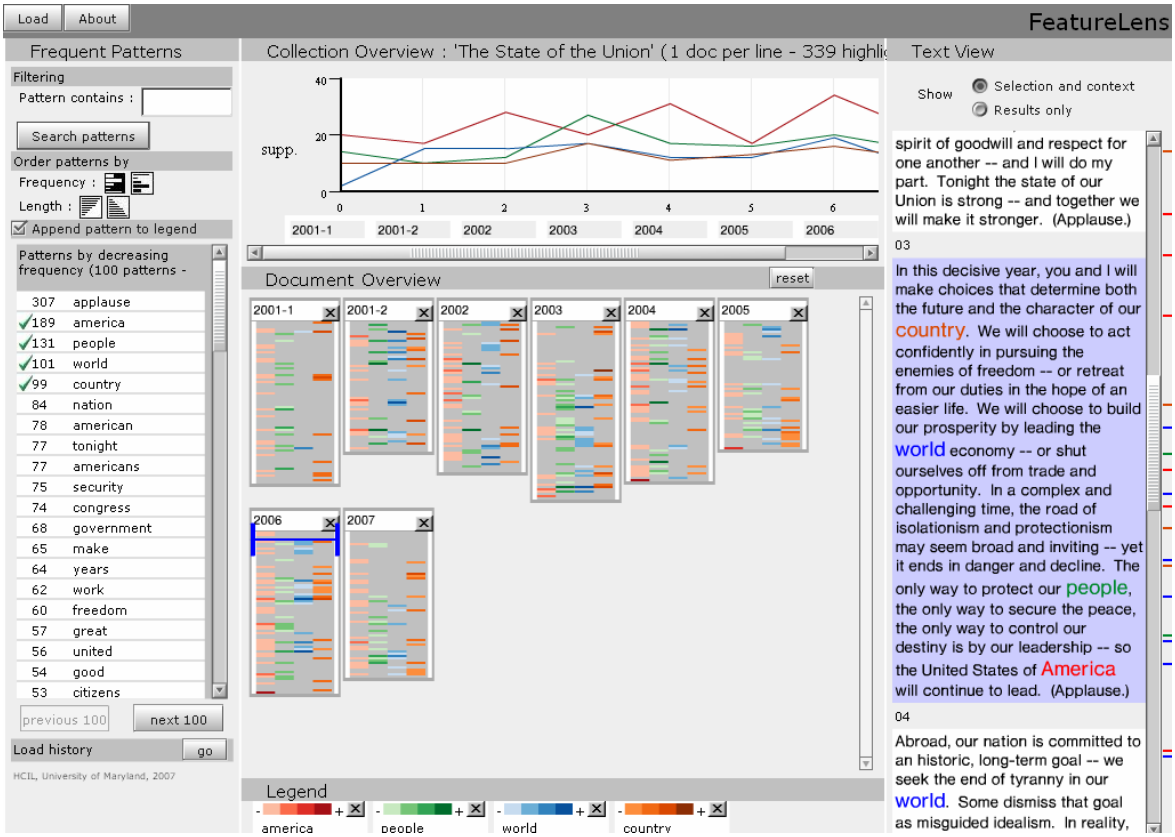


Figure 3.32: The FeatureLens application used with the State of the Union data set [32].

Instead of using pre-defined patterns, FeatureLens [32] relies on visualization to convey patterns mined from text collections. In FeatureLens, two types of patterns are used: frequent words and frequent trigram item sets [32]. Each document is represented by a compact rectangular area, in a similar fashion to the SeeSoft visualization. Each pattern is color-coded to highlight its appearances in the text, with a more frequent pattern being assigned a darker color [32]. Figure 3.32 shows an example analysis using FeatureLens on the State of the

Union speeches, whereby we can see co-occurrences of different patterns within documents. In FeatureLens, it is easy to see the co-occurrences of patterns within a text. However, to see the contexts in which a pattern is matched, users still need to read the full text of pattern-bearing paragraphs.

The POSVis [148] application is designed to support digital humanities researchers in studying characters in a novel based on part-of-speech information. Figure 3.33 shows POSVis being used to analyze the book *The making of Americans*, in which named entities (listed in the upper left side) can be used to explore the text. Given a selected named entity, words within a defined vicinity in the text are visualized in a word cloud or a network, and they can be filtered by part of speech. POSVis also enables comparisons of usage contexts by using a word cloud, which encodes frequencies of occurrence of surrounding terms in one section using font size and those in another section using color (c.f. Figure 4 in [148]). POSVis, however, does not focus on the semantics of surrounding words, as it is skewed toward analysis and interactions based on syntactic information.

Though not intended for concordance analysis, the SeeSoft visualization [35] can intuitively shows the locations and contexts of occurrences of terms in documents, which are represented by vertical bars. In fact, the SeeSoft developers experimented with applying it to the display of text to highlight where characters appear within a book, and which passages of the Bible contain references to particular people and items [59]. The SeeSoft visualization is also adapted and modified in a visualization by the New York Times[8], as shown in Figure 3.34, which shows how words are used in context in the State of the Union speeches given by a US president from 2001 to 2007. In this visualization, each document is depicted by a set of stacked bars representing its paragraphs, and the distribution of a seed term can be seen via a set of dots representing a term's occurrences in the text. When users click on a dot in the visualization, they are presented with the term-bearing paragraph, with the term's appearances highlighted in the text viewer. While this visualization appears to be simple and easily comprehensible, users are still left with a lot of text to analyze. Although this is a reasonable action to take in some cases that require deep reading, it is not easy for users to

[8] `http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html` (Last accessed Jan 12, 2011)

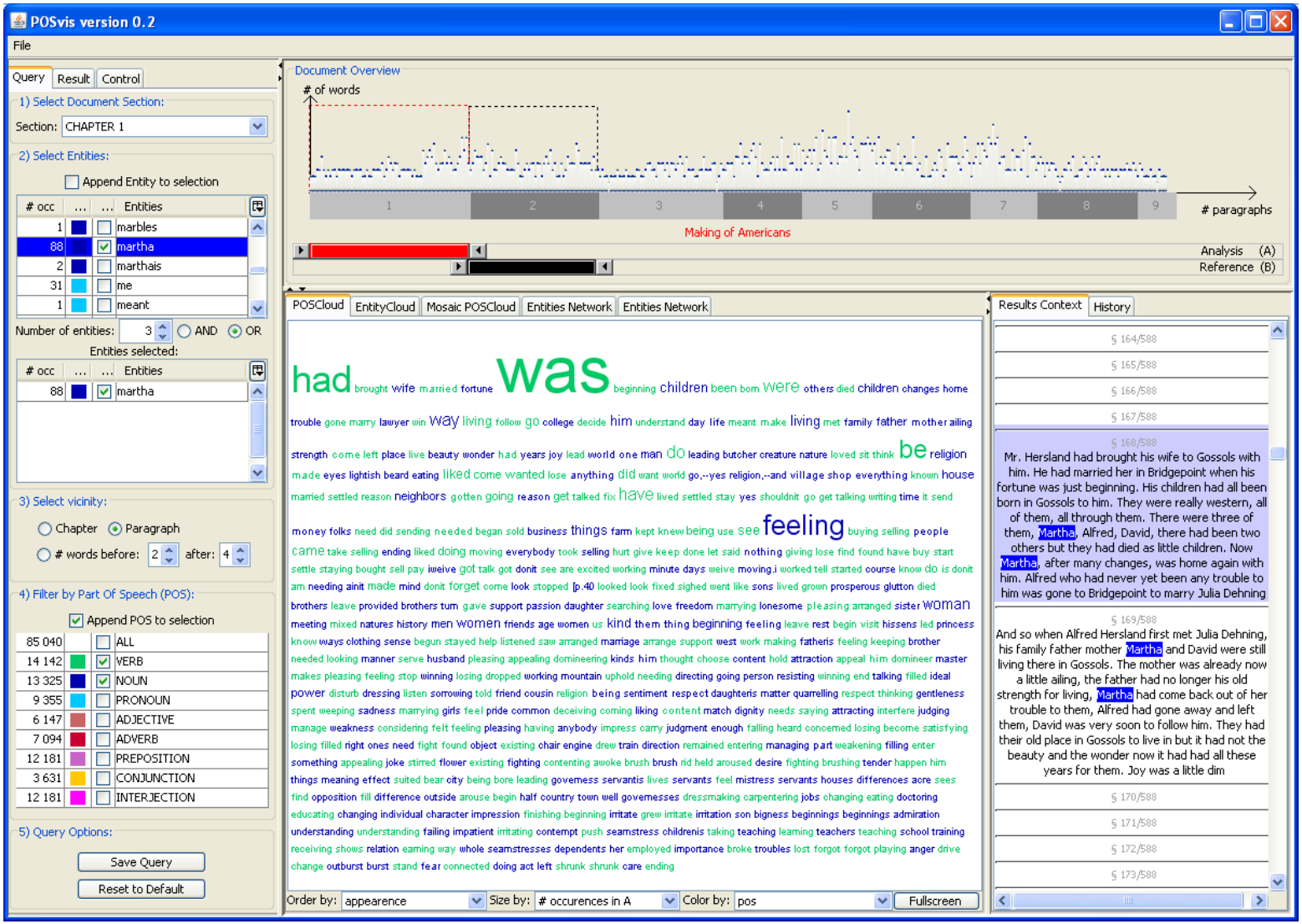


Figure 3.33: A POSVis visualization [148].

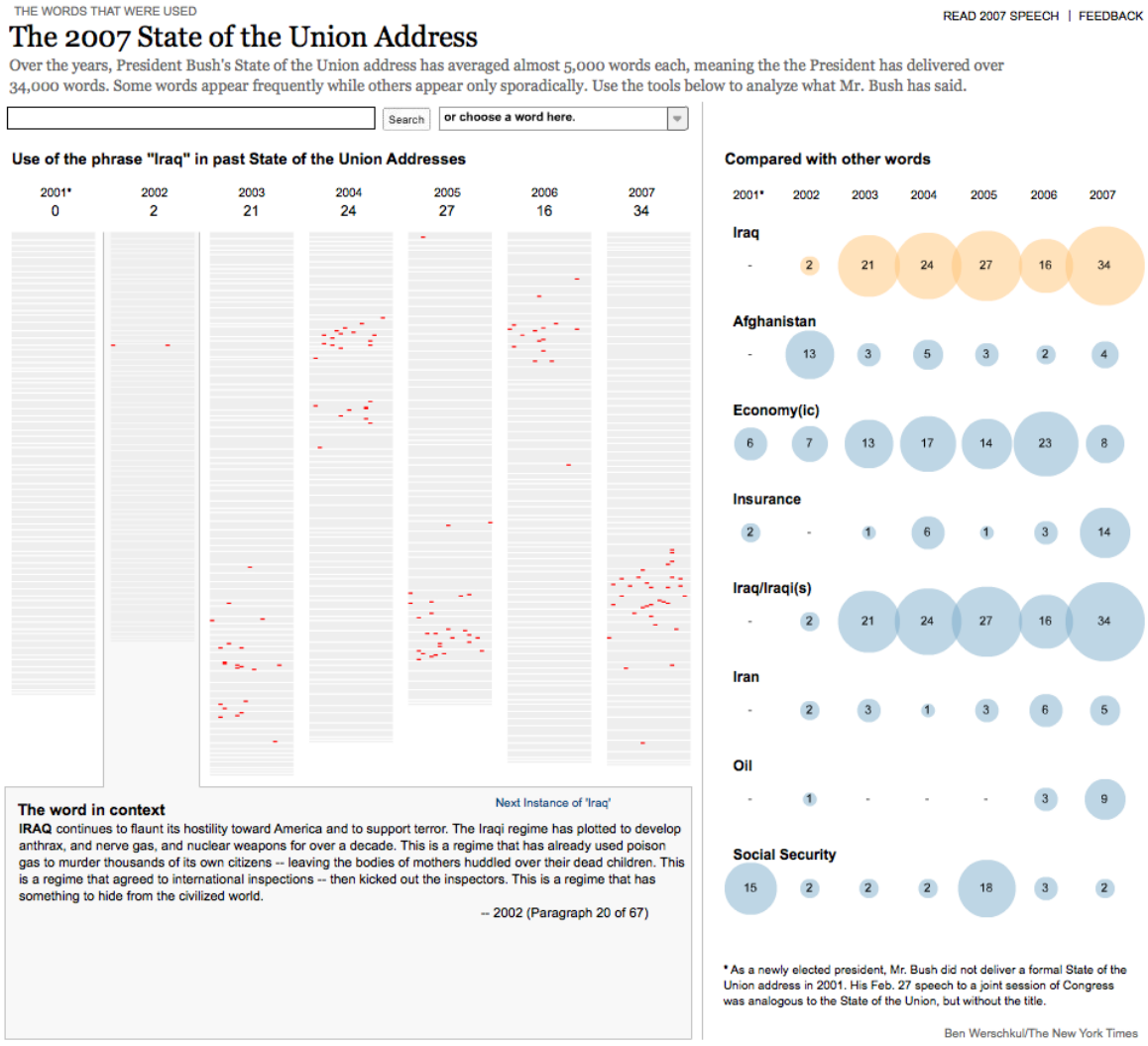


Figure 3.34: A New York Times visualization of a term's usage contexts within State of the Union speeches.

quickly grasp the usage contexts of a term in a document or compare its surrounding contexts in one document versus another.

In sum, there are many visualizations developed to support users in visual concordance analysis tasks, ranging from a simple tabular form to more sophisticated graphs. These visual means all aim at showing how a word is used or how it is related to another. Some visualizations solely focus on the relationships between a word and its context (such as Word Tree and Phrase Nets), others also keep users informed about its distribution (such as FeatureLens and the New York Times visualization). More importantly, as the number of words in a long text can be large, visual scalability is an inherent challenge in most of these works.

3.6 Visual Analysis of Trends in Text

Temporal properties of most data types are of significant importance in understanding the evolving dynamics of the domain of interest. Textual data type is no exception to this phenomenon. In order to gain an understanding of what kind of changes manifest over time based on textual content of documents, temporal-based analysis of document is an important area of research and development. Applications of trend analysis of document collections are vast, such as tracking technology advances based on patents and publications, or performance of a listed company via its regularly published business reports, just to name a few. In this section, we highlight some innovative works in this area.

ThemeRiver [53] is a visualization that is based on the river metaphor to depict thematic changes within a collection of documents. In ThemeRiver, each theme is shown as a current bound by continuous curves within a river flowing from left to right along a timeline below it. As such, the evolution of various themes and their strengths within a text collection over time is reflected by the composition and changing widths (vertical distances) of the color-coded currents within the river visualization [53]. Figure 3.35 shows an example ThemeRiver visualization of a collection of news articles from July and early August 1990. A current shrinks or bulges to depict a decrement or an increment of a theme's strength at a point in time [53]. There are also markers to highlight related historical events along the top. While the river metaphor lends itself nicely into depicting the evolving dynamics of the

content within documents over time, it has certain limitations. As temporal data are not continuous, the ThemeRiver representation contradicts this fact due to the interpolation of values to obtain trend curves. In fact, one of the subjects participating in a formative study of ThemeRiver made a point that she was not sure exactly of what the data values meant [53]. In addition, it is a known concern of stacked bar graphs (a category that ThemeRiver belongs to) that they "*are problematic because the shape of any given line is determined in part by the shapes of the lines below it, thus potentially misleading the interpretation of the graph's values*" [59]. Lastly, visual scalability is also an issue here, as only a limited number of theme currents can be displayed, otherwise, they might become unwieldy for users.

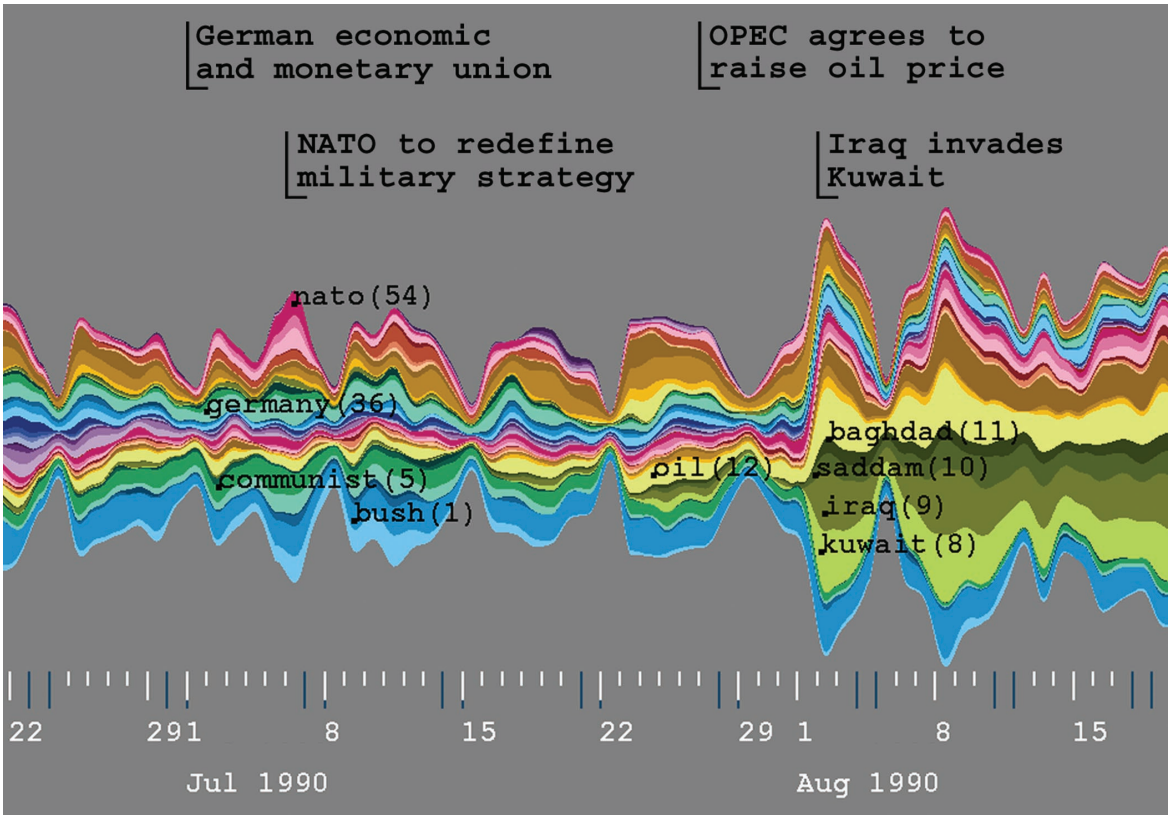


Figure 3.35: A example ThemeRiver visualization of news articles [53].

Another more recent work to be noted is TIARA [83], which uses the ThemeRiver metaphor to show the ebbs and flows of topics within a collection of documents over time. Unlike ThemeRiver, TIARA employs an advanced text analysis technique called Latent Dirichlet Allocation [8] to obtain a topic model of a collection of text. These topics are then

visualized in a similar way to themes in ThemeRiver, however, they are better annotated with terms belonging to topics at different points in time. Therefore, TIARA can provide a more informative visual summary of a collection of text (see Figure 3.36). Similar to ThemeRiver, TIARA only focuses on visual analysis at collection level, without any support for further exploration at document level except for text highlighting.

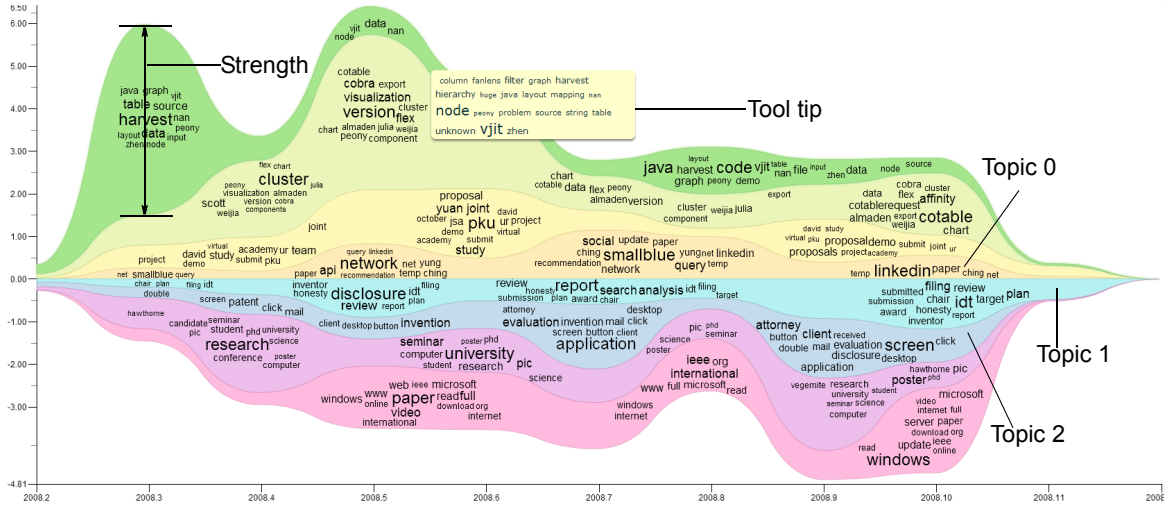


Figure 3.36: A visual summary of 10,000 emails produced by the TIARA system [83].

Another particular area of research in visual trend analysis of document collections is concerned with the editing history of documents. This kind of research is particularly helpful in domains where collaborations between multiple people are at the core of their tasks. Many research efforts in this area focus on histories of Wikipedia articles as they are made publicly available.

history flow [145] is a visualization technique to depict the collaborative dynamics of how documents are created and modified and to make trends in revision histories immediately apparent. Figure 3.37 shows a *history flow* visualization of the editing history of the Wikipedia article on "*Microsoft*". In a *history flow*, different versions of a document is represented as vertical bars placed in parallel. Within each bar, authorships of all sections are color-coded accordingly. To visually illustrate how different versions relate, shaded connections are used to link sections of text that have been left untouched between consecutive versions. Vertical bars can also be spaced unevenly to indicate temporal distances between different versions.

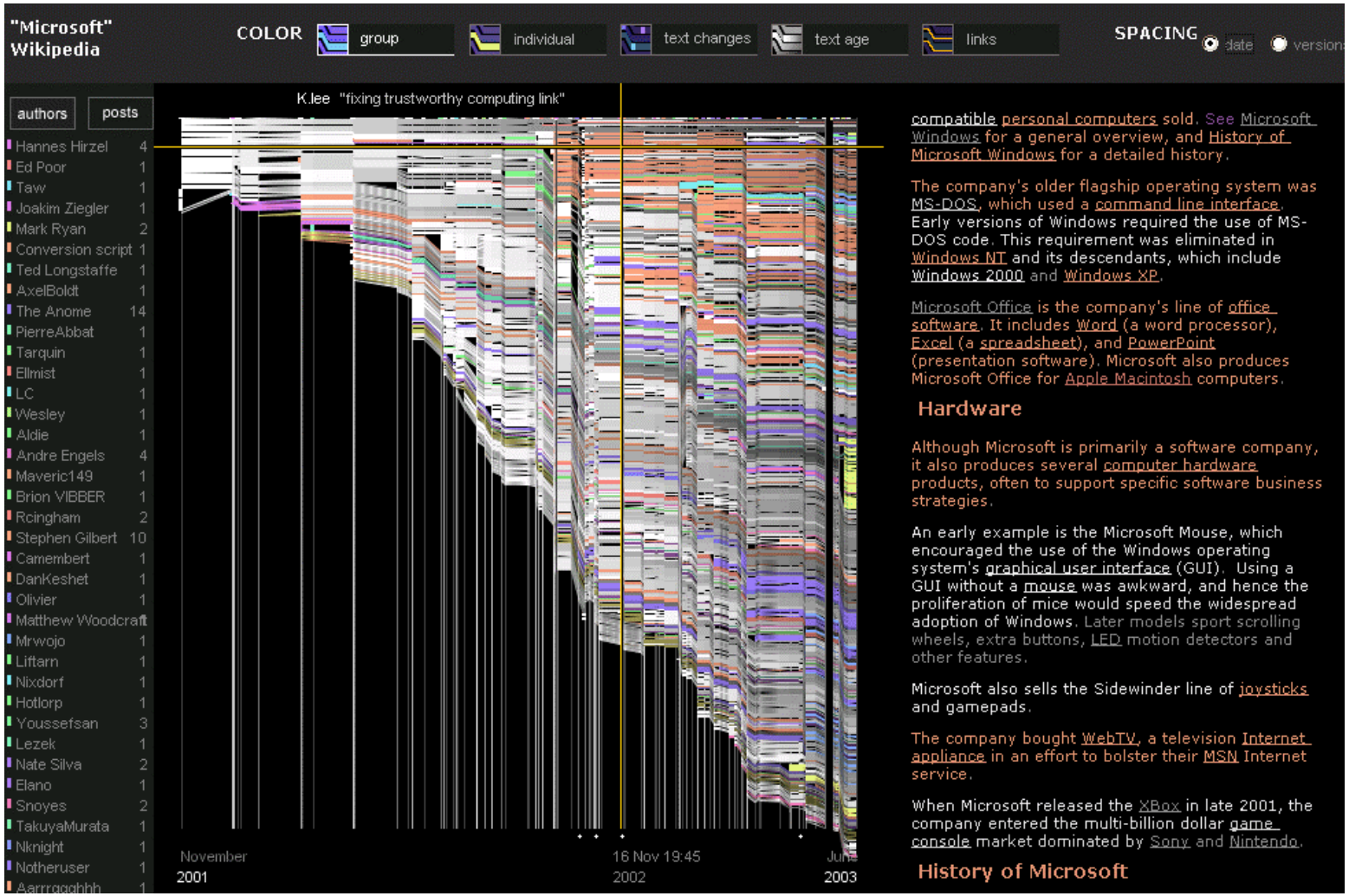


Figure 3.37: A *history flow* visualization of the editing history on the Wikipedia article on “Microsoft” [145].

As seen in Figure 3.37, the Wikipedia article on "*Microsoft*" was edited quite frequently by many collaborators and its content has a near-constant growth [145]. Even though *history flow* only focuses on a document at a time, it seems to be effective in depicting how parts of a document evolve in relation to each other's position and author.

There are also many other efforts in this area, we briefly highlight some of them here. Chromograms [152] are simple and compact color-coded bars used to encode the editing histories for a collection of pages. Chromograms can be used to characterize the activity trends of highly active contributors, such as their styles and rhythms. Figure 3.38 shows a screenshot of the Chromograms application, in which a timeline view shows different types of edit made on a page over time.

Figure 3.38: The Chromograms application with a timeline view of the editing history of a Wikipedia article [152].

WikiDashboard [134] is a visual overlay for Wikipedia pages to show a summary of who

edits how many revisions on each page. As shown in Figure 3.39, this embedded visualization provides users with a convenient way to be aware of editing activities and trend on the page.

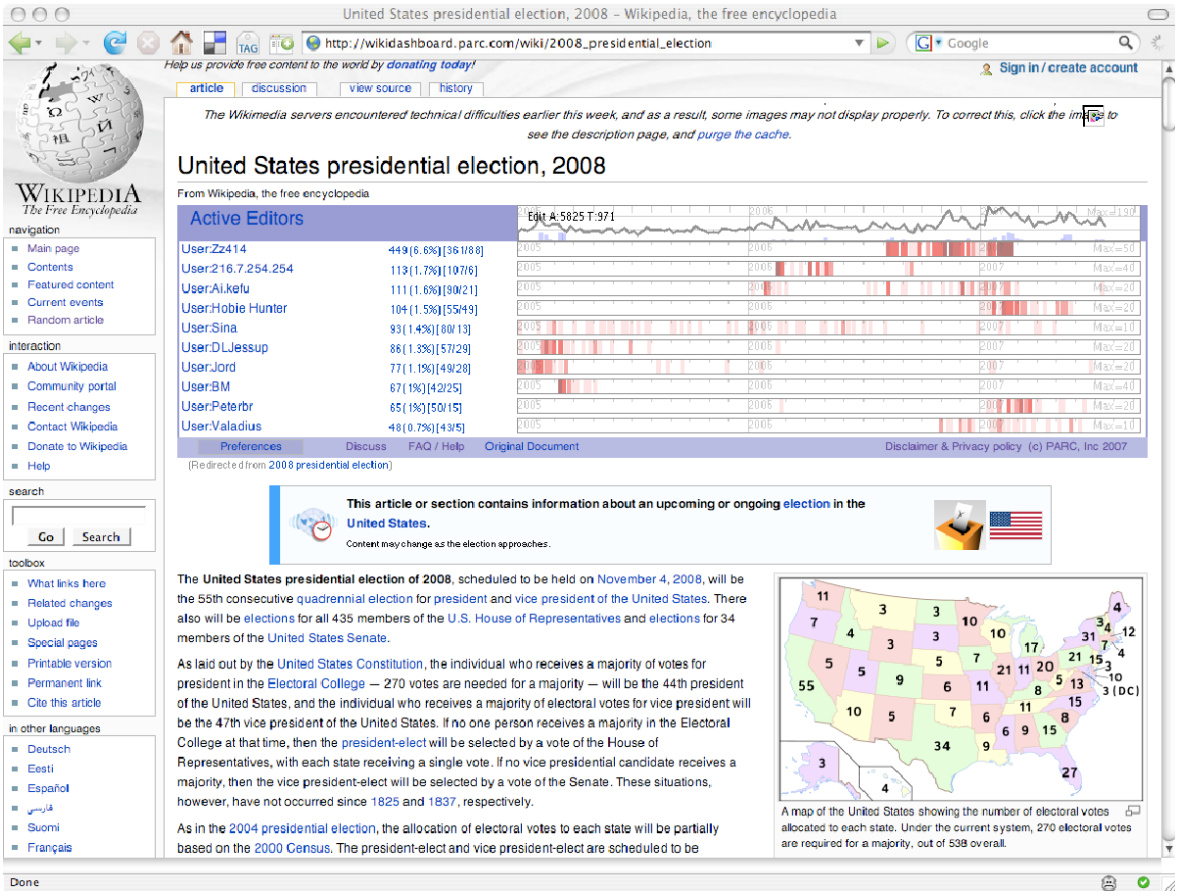


Figure 3.39: A WikiDashboard of the editing history of a Wikipedia article [134].

A recent addition is the iChase visualization [106], which focuses not just on visualizing the editing history of a single page, it attempts to show the relationships between multiple articles, authors and time period at once. As shown in Figure 3.40, it uses a combination of focus+context matrix displays (in a manner similar to the TableLens visualization [103]) to show the relationships between articles and their edit timestamps, as well as the correspondence between contributors and articles edited by them, and a timeline to show the evolution of the number of articles and contributors active over time.

As diverse as they are, all of the approaches highlighted in this section employ visual means on top of a horizontal timeline, as this is traditionally expected by users when seeing

Figure 3.40: An iChase visualization of one week activity on the WikiProject "*Louvre Paintings*" [106].

time-related presentations. Apart from many efforts on visualizing trends of editing histories of wikipedia articles, innovative approaches, such as that of TIARA, have recently started focusing on more in-depth analysis of documents' contents using text mining techniques, in order to bring more detailed aspects of text to trend visualizations.

3.7 Visual Analysis of Text Streams

The variety of works highlighted thus far mainly focus on either corpus level or document level exploration of text. They treat a collection of documents statically, i.e. making an assumption that the set of documents does not change over time. However, many types of textual data do come in sequence, for instance, emails, newswires, and more recently, social media content such as tweets. Hence in many cases there is a need to analyze dynamic streams of text. Existing techniques such as those discussed in Section 3.1 are not suitable for this task because they focus only at corpus level, and hence "*stories that span a week may be overshadowed by major news trends that span a year*" [110]. In addition, since content of text streams changes constantly, corpus level methods need to be re-executed to get updated when new data come in [110]. These constant updates would inevitably cause performance

issues with interactive systems. In what follows, we highlight some key approaches toward visualizing text streams.

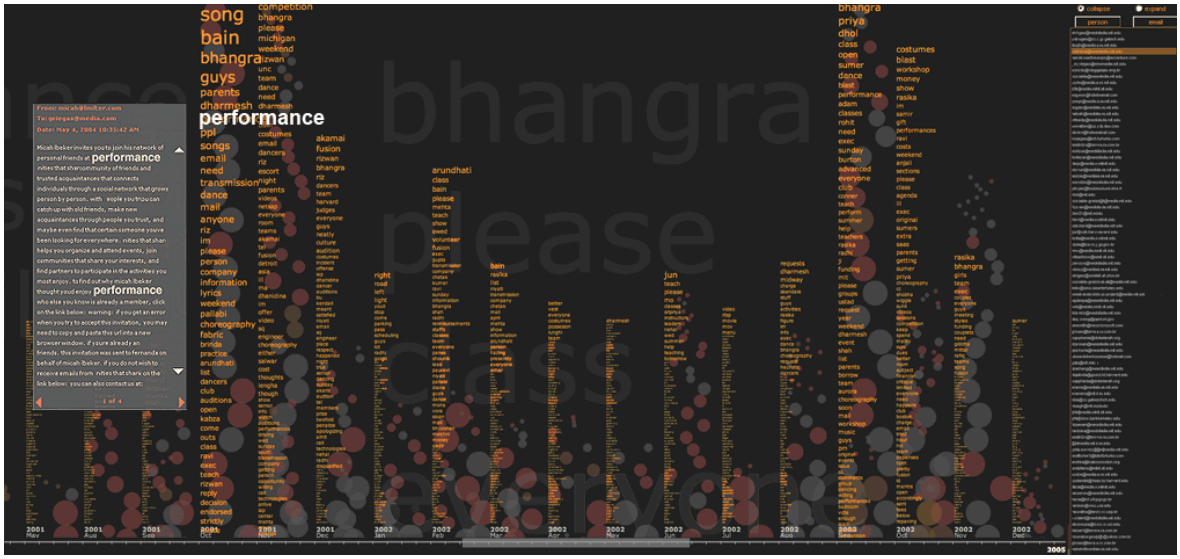


Figure 3.41: A Themail visualization showing a user's email exchange over 18 months [143].

For streams of emails, Themail [143] is a typographic visualization of email exchange histories. This visualization is worth mentioning here because unlike other works in email visualization which focus only on the header information (senders, recipients, dates, etc.) to support social network analysis, Themail utilizes the content of email messages to bring about a deeper understanding of how a user's email exchanges evolve over time. As shown in Figure 3.41, Themail consists of columns of keywords placed along a timeline. The *yearly keywords* are the most used terms over a year of email exchange and are shown as large faint words in the background. The *monthly words* are frequently used terms in a month, and are shown in yellow in the foreground in different sizes depending on their frequency and distinctiveness. Stacked colored circles representing outgoing and incoming messages within a month. Interactions are possible on *monthly words* to bring up details of emails containing them [143]. While Themail can provide users with a visually appealing depiction of the overall themes of a user's email communications as well as more detail on demand, there are certain limitations with this visual text analysis approach. Firstly, sometimes it is not straightforward to associate the content in an email account to a its owner (for instance, when an email is sent on someone's else behalf), which might lead to misleading interpretation

[143]. In addition, without any visual abstraction mechanism, it might not be easy to grasp the overall topics of communications within months having very busy traffic.

It is also worth noting the story flow approach [110], which focuses on identifying and illustrating the context that relates new information with old one from an information space that is continuously changing. This can be particularly helpful in analyzing newswires stories / events that span multiple documents and time intervals. Within a time interval, key words are extracted from each document published or received within that period, and all of them are then grouped into coherent themes based on the similarity of their document associations [110]. Each document occurring within the past n intervals is assigned the theme for which it has the highest total score [110]. Each theme is labeled with a term that best represents documents assigned to that theme. These themes are used to describe developments of stories in a story flow visualization, as shown in Figure 3.42. As stories may intersect, split and merge, documents assigned to a theme within one interval might be assigned to a different theme in the next interval and there are additions of new documents to an existing story as well as aging out of documents older than n intervals [110]. While this approach makes large stories standing out from the rest, it is not that easy to spot the evolution of stories that span a shorter period of time. In addition, at any point in time, a document is only assigned to a theme, which is not always the case in practice.

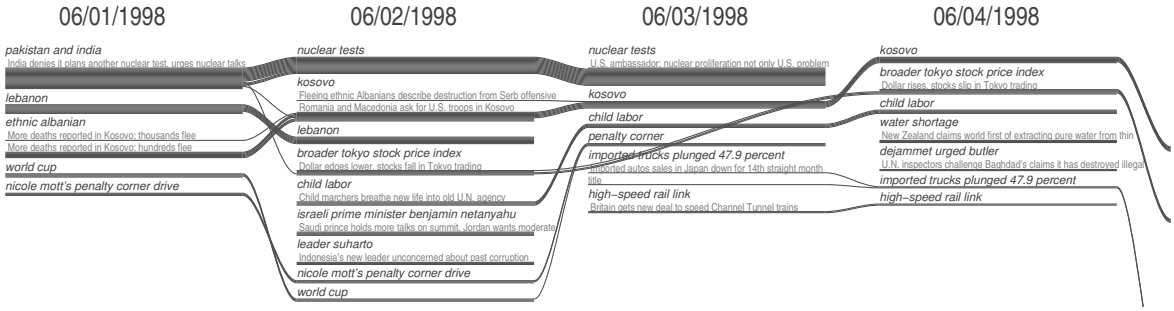


Figure 3.42: A story flow visualization showing the development of different themes within newswires stories [110].

An important and more recent kind of text streams comes from social media such as tweets from Twitter's microblogging service and status updates from social networking platforms. The major differences between this kind of data and other textual data are that they

tend to be really short (limited to 140 characters in the case of tweets), they are usually not considered as contained in documents, and there is a lot of noise in terms of vocabulary used such as abbreviations, texting languages. As a result, these peculiarities inherent to social media streams necessitate innovative research to help users in casual browsing or analyzing a large amount of social data.

Figure 3.43: An Eddi visualization showing major topical themes and related tweets [4].

Eddi [4] is topic-based interface designed to support users in browsing their Twitter streams. Eddi attempts to help users in coping with being inundated with a tremendous amount of tweets in their Twitter feeds by identifying topics and then grouping tweets according to those topics. Eddi employs the Stanford part-of-speech tagger to extract noun phrases from tweets, and use them to query for web documents from an external search engine [4]. Eddi then retrieves the popular terms in those documents and ranks them as topics [4]. These topics are used to support users in browsing their Twitter feeds, as shown in Figure 3.43. In this interface, a tag cloud is used to show the "*major topical themes*", with the size of a topic proportional to the number of tweets in that topic, and brushing interactions allow for a quick view of top tweets in a topic. A topic dashboard is used to show tweets

that users might be interested in, based on what they themselves tweet about [4]. Eddi also provides a timeline showing trending topics within the past week (c.f. [4] for examples). The authors themselves noted one of the key weaknesses of the approach proposed in Eddi is that it might not be able to get the right topic for tweets containing figures of speech. In these cases, the returned documents from a web search might not lead to those containing the right intentions or meanings of the original tweets.

3.8 Other Notable Approaches

There are also a number of noteworthy works in text visualization that do not fall into any of the categories discussed thus far.

The visual analytics system Jigsaw [126] aims at supporting analysts in investigating relationships between entities across documents. In Jigsaw, entities such as places, persons, organizations are extracted from documents by the ANNIE information extraction component within GATE [26]. ANNIE relies on a type of lexical resources, which are gazetteer lists of entity names used for annotating the text. In Jigsaw, entities that appear together in at least one document are considered to be connected [126]. Multiple coordinated views such as list view, network view, etc. are used to show the relationship between entities and documents. Figure 3.44 shows a list view within Jigsaw in which each list consists of entities of a certain type and links between entities indicate whether they appear together in some documents. The links' highlighting colors encode the number of documents contain both connected entities [126]. Users are also provided with other views to explore from different perspectives. However, visual scalability seems to be an issue in Jigsaw, as the authors of Jigsaw remarked themselves that Jigsaw was designed toward short documents that have "*about 1-6 paragraphs*" and that it became "*less useful as the document size increases because the higher number of entities per document swamps the display*" [126].

Another work to be noted is DocuBurst [22], which is a visually appealing application for knowledge-based text analysis. As shown in Figure 3.45, DocuBurst uses a radial space-filling layout, in a similar fashion to SunBurst [127], to display frequencies of words in a document that are related in meaning to the seed term. The analysis is based on a language

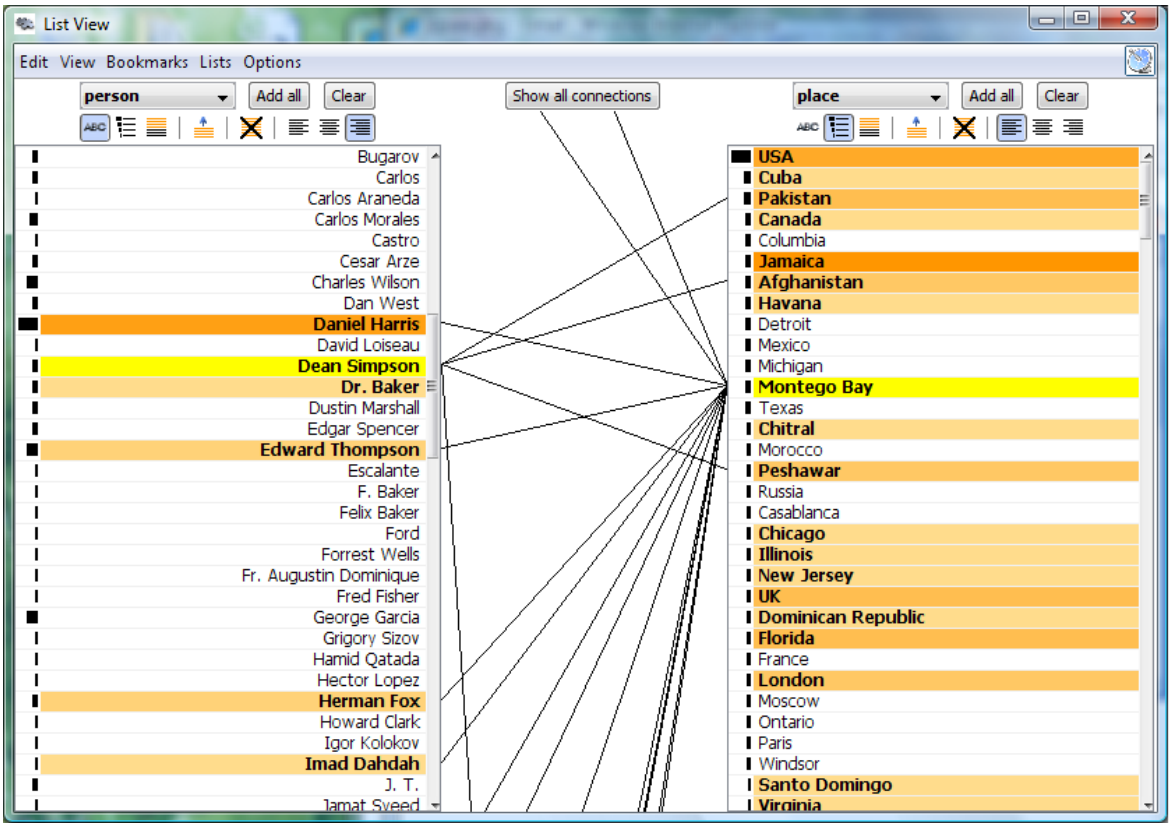


Figure 3.44: Jigsaw's list view of connections between entities within documents. Selected entities are highlighted in yellow. Entities that are connected to others are highlighted in orange [126].

structure called WordNet [39], where the hyponymy ("*is-a*") relationships between words are obtained for visualization [22]. In DocuBurst, the seed term is placed in a center circle, and its related words are recursively placed at the wedges in the next level. An annex of DocuBurst shows contexts for concordance analysis in the same manner as the Concordance software, i.e. in a tabular form with the seed term in between and the surrounding text on the sides. The authors of DocuBurst themselves note that text analysis based on WordNet as a lexical resource is not without limitations, as its "*sense-divisions are too fine-grained for many computational applications*" and hence some sort of semantic abstraction is necessary. Finally, because of the amount of space that a DocuBurst visualization requires, it tends to be more useful for detailed analysis of a long document on its own than for comparing a large number of documents at once.

Wordle[9] [146] is a popular web-based application to create word clouds from text. Wordle is different from other typical tag-cloud like displays in that it focuses heavily on aesthetic aspects such as typography, color, and composition [146]. While other approaches to generate tag clouds place glyphs in such a way that their bounding boxes are separated, in Wordle small glyphs can be placed within larger ones, as long as they do not intersect [146]. Moreover, Wordle also lets users have a say in how the resulting visualizations should look like by setting various parameters, for instance, layouts, color themes or typefaces, as well as in how they want to share the resulting Wordle visualizations. A study on how Wordle is used in the wild has shown that this application actually was predominantly used for various non-analytical purposes, ranging from education, mainstream media to personal communications [146]. In this non-traditional context, Wordle's success has emphasized the importance of enabling users to engage in creating aesthetically pleasing and expressive visual artifacts instead of just following conventional principles of visualization design [146].

Parallel Tag Clouds [23] is a visual text analytics tool designed to support users in analyzing the differences in terms used in faceted text corpora. In this approach, a statistic indicating the probability that the frequency of a word in a corpus differs significantly from another is calculated and only terms having values above a threshold are retained for visualization (c.f. [23] for technical details). Visually, a Parallel Tag Cloud is a combination of

[9] `http://www.wordle.net/` (Last accessed 31 August 2011)

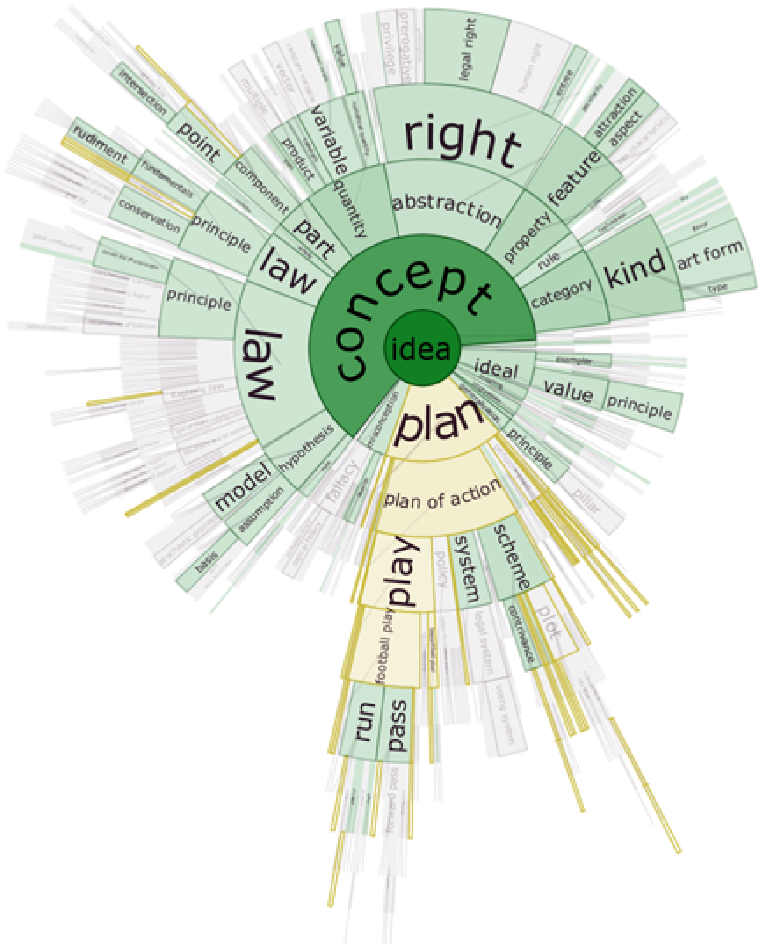


Figure 3.45: A DocuBurst visualization of a science textbook, with the seed term “*idea*” [22]. Highlighted in yellow are branches that contain words having the characters ‘*pl*’.

Figure 3.46: An example Wordle [146].

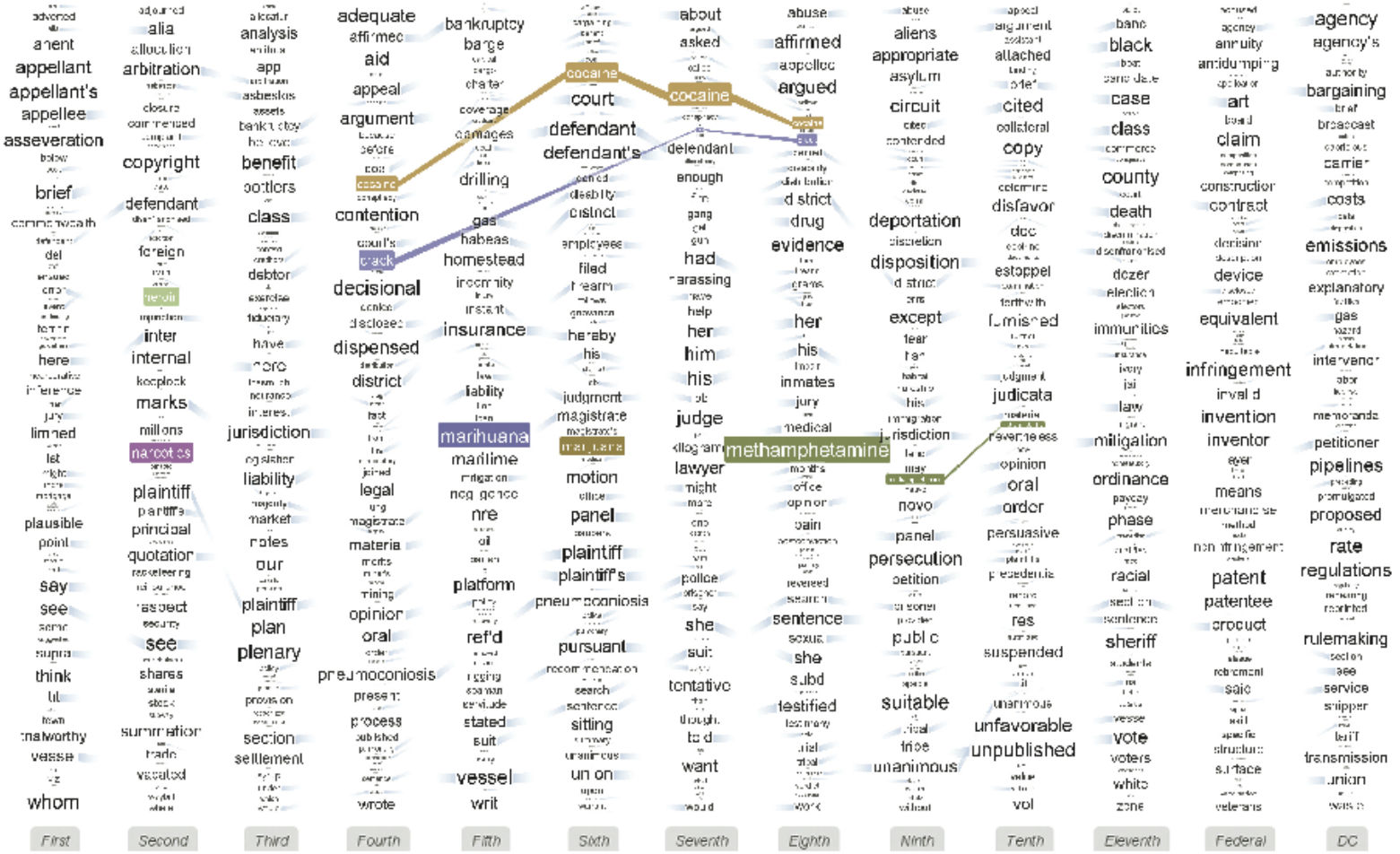


Figure 3.47: A Parallel Tag Cloud of written decisions of the US Circuit Courts of Appeal [23].

parallel coordinates [65] and a font-size based tag cloud. As shown in Figure 3.47, selected terms are alphabetically sorted and put into columns representing subsets of the corpus, and edges connect the same terms across different subsets. Figure 3.47 illustrates how different types of drugs are mentioned among the circuits. In addition, a Parallel Tag Cloud can also be useful to reveal a significant absence or presence of a word [23].

3.9 Summary

Text is inherently categorical and its multi-dimensional nature makes it challenging for analysts and casual users alike to gain an understanding from a large amount of documents. This challenge has motivated a lot of research and development on techniques and methods to support users in exploring and understanding from collections of unstructured text sources. In this chapter, we have discussed a variety of interactive tools supporting visual exploration of text collections. The breadth and depth of these existing tools speak volumes for the significant role that they play in reducing the cognitive load required to obtain information contained in text.

In the next part, we will move on to present our research contributions in this area.

Part III

Core

Chapter 4

Ontology-based Visual Exploration of Text Collections

As discussed in the previous chapters, information visualization is an essential mechanism to support users in exploring various facets of complex information spaces where non-visual information needs to be visualized and interacted with, such as text collections. Given the amount and the unstructured or weakly structured nature of text documents that scientists and analysts alike have to deal with, text visualization applications can help users gain the needed understanding or insights in a timely manner. Depending on the application domains, insights obtained from the exploration and analysis of text collections can enable users to understand, for instance, the distribution of topics or to identify trends and linkages between different entities [117].

While many existing tools provide support toward visual exploration of text collections in different ways as highlighted in Chapter 3, the majority of them present findings that are independent of users' interests and knowledge. This setback is an important one in many cases whereby analysts, in following their business practices or expertise, would focus on a very specific set of pre-defined entities as well as utilize their knowledge about such entities (different linguistic variations of their mentions within a text, such as abbreviations, or phrases that mean the same thing). Consequently, a visual presentation of entities that are outside of analysts' spheres of interest could be counterproductive to the analytical tasks at hand. Therefore, when a specific set of pre-defined entities are of significant importance to

users' exploratory goals, it is essential that the visual exploration process can be aligned with users' interests and knowledge. Interested readers can refer back to Section 1.1 in which we discuss an example use case in the business analysis domain where this need is particularly relevant. This kind of situation has motivated research to make exploration of text collections more tailored to specific needs of users.

While in the discussion of existing tools in Chapter 3 we have mentioned their strengths and weaknesses individually, here we briefly summarize some of the gaps present in these works in terms of how they can support users in incorporating their interests and knowledge into the visual exploration process before going into our proposed approach in the next section. To facilitate comparisons, this gap analysis is also done relatively to Shneiderman's visual information seeking mantra "*Overview first, zoom and filter, then details-on-demand*" [120] that we adopt in our own work.

In terms of supporting users in gaining an *overview* of a text collection, the approaches reported in Section 3.1 rely on dimensionality reduction methods to derive a representation of documents in a 2 or 3-dimensional space. This space is derived in such a way that documents that are close to each other in the original multi-dimensional space remain so in the newly derived space. Among these approaches' key limitations is the difficulty for users to interpret the newly generated dimensions. Furthermore, these methods only enable exploring inter-document similarities without taking users' interests or knowledge into account. Meanwhile, the approaches reported in Section 3.2 can provide users with an overview of a text collection based on a knowledge structure such as a taxonomy or a domain ontology. Among them, DocCube [90] stands out as providing an overview of the dispersion of documents along various dimensions and it also makes use of the hierarchical structure within an ontology to let users explore at different levels of aggregation. The 3D visualization technique used in DocCube, however, suffers from the common problem of occlusion. In sum, most existing techniques in this area appear to suffer from usability issues whereby understanding of the relationships between items shown in an overview display is not very well facilitated.

In terms of supporting users in *filtering* tasks, a number of query terms based techniques enable users to see how many documents meet each possible combination of the query terms used for filtering. These techniques, however, are limited in their own layout algorithms in

certain cases (e.g., ambiguous interpretation at intersections within the VIBE display [95] and complex visual encodings in the case of InfoCrystal [125]). Moreover, the lack of a reusable conceptual structure such as an ontology that allows for organizing query terms of interest to users into a hierarchical structure means that exploring at different levels of abstraction is not feasible with such systems. Apart from those query terms based filtering techniques, many applications support faceted filtering. Most of them, however, require clean and accurate metadata describing the classifications of resources into various facets. This has a number of known limitations: (1) clean and accurate metadata are not always available for large resource collections, (2) if there is a mismatch between users' mental models and the set of facets being used for filtering, it would result in negative effects on user experience. Moreover, when the resources being filtered are text documents, existing faceted browsers only treat the relationships between documents and facet values as binary ones, in a similar manner to other non-content bearing resources, and hence the contents of documents are not taken into account.

With respect to supporting users in investigating the *details* of documents on demand, there are two main areas of research: (1) visual analysis of document structures and term distributions and (2) visual concordance analysis. For the former, many techniques rely on the linear structure of documents to show where terms appear within them using intuitive layouts (mostly involving blocks or bars representing chunks of text within a document placed linearly next to each other either horizontally or vertically). However, these techniques are not without limitations when it comes to visualizing the distributions of a large number of entities (e.g., TileBars [56]) or at a finer level of granularity (e.g., Seesoft [37]). For the latter, there is a diverse set of tools and techniques to assist users in understanding the usage contexts of terms within documents. These tools exist in various forms, from a tabular representation to a large graph. A common limitation in these techniques is with regard to the amount of text that still needs to be shown so that users can understand the usage contexts of terms. This setback can become a significant challenge for users when a document is large and many parts of it contain a term of interest to users.

In this chapter, we discuss our proposed approach and the research prototype, which also paves the way for further research in Chapters 5, 6 and 7.

4.1 Proposed Approach

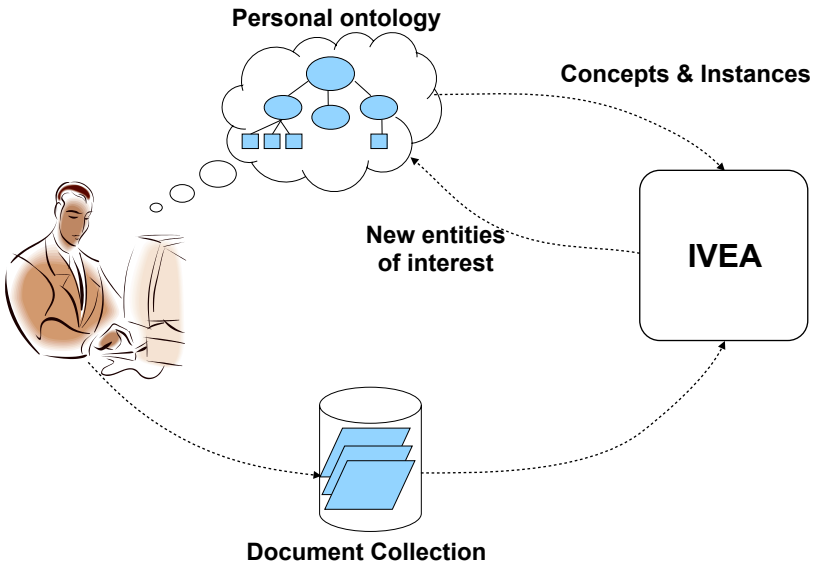


Figure 4.1: Approach Overview

In addressing the limitations outlined in the previous section, we proposed an approach, as illustrated in Figure 4.1, to help users focus on aspects of a text collection that they are particularly concerned with by:

- Involving users at an early stage in the exploration process, whereby they define their spheres of interest by encoding the important entities and their relationships into a personal ontology. For each instance in this ontology, users can also define a list of different linguistic variations associated with it so that its mentions in the text can be identified.

- Leveraging upon the hierarchical structure of entities defined in the above-mentioned ontology to allow users to explore various dimensions of a text collection at different levels of detail.

- Employing coordinated multiple views to allow users to look at documents in a text collection from different perspectives. These views are designed in such a way that they can facilitate the interactions outlined in Shneiderman's visual information seeking mantra.

- Suggesting users with frequent phrases within documents to keep the personal ontology updated with new entities potentially matching their evolving interests. With the newly added entities, the personal ontology becomes a richer and better representation of their interests and hence can lead to more closely tailored subsequent exploration and analysis experiences. This loop of defining and enriching an ontology containing entities of interest is similar to the metaphor of "*Plant a seed and watch it grow*".

In comparison with previous works on query terms or points of interest based visualizations, such as VIBE [95] and InfoCrystal [125], while they can allow users to focus on documents containing entities of interest, they do not take advantage of any re-usable knowledge representation. As a result, users with knowledge about certain entities will have to encode that knowledge into more complex queries, and this time-consuming query formulation process needs to be repeated every time another document is analyzed. In our ontology-based approach, a personal ontology can be defined by users once, and subsequently used many times. Furthermore, our approach also enables users to explore a text collection at different levels of detail based on the hierarchical relationships between different entities. Our approach is also different from other works on knowledge-based visual text analysis in a number of significant ways. First, in our work there is a loop of knowledge utilization and knowledge acquisition whereby new entities of interest is incorporated over time. Second, in our work, by adopting Shneiderman's mantra, we provide users with a more complete set of interactions that let users see an overview, filter out uninteresting documents, and investigate more details of documents on an individual basis. Third, our work also comprises of contributions that go into much more depth of visual text analysis, including: a visualization that supports abstraction and ordering of items in faceted filtering, an effort in tackling the visual complexity of TileBars that can benefit visual analysis of entities distribution within documents, as well as a topic-based content abstraction approach for visual concordance analysis. These contributions address many limitations summarized at the beginning of this chapter and will be covered in detail in later chapters.

It is also worth noting that while any personal ontology can be used within the proposed approach, we initially employed an ontology that was developed within the Gnowsis project and based upon the *Personal Information Model* (PIMO) [113]. Since the PIMO ontology

acts as a formal representation of the structures and concepts within a knowledge worker's mental model [113], it can be employed as a means to integrate personal interests and knowledge into an exploratory visualization tool. The PIMO ontology comes pre-loaded with some initial concepts that might provide a starting point for users. This is useful as many knowledge-based systems suffer from the cold-start problem, in which systems require initial user inputs to be able to produce valuable outputs, but without seeing any meaningful outcomes users are less motivated to make an effort to give such inputs.

The proposed approach is initially realized by a visualization prototype called IVEA, which leverages upon the PIMO ontology and the Multiple Coordinated Views technique. IVEA allows for an interactive and user-controlled exploration process in which the knowledge workers can gain a rapid understanding about a text collection via intuitive visual displays. The design of IVEA's initial visual interface and interactions is described next.

4.1.1 Visual Interface

We employ the Multiple Coordinated Views technique [109] to facilitate interactions suggested by Shneiderman's mantra in designing IVEA. Its interface is shown in Figure 4.2 when used to explore a collection of documents via an example pre-defined ontology. The use of the Multiple Coordinated Views technique enables users to understand a text collection better via different visual displays highlighting various aspects of it, as well as via the interactions with and coordinations between those displays.

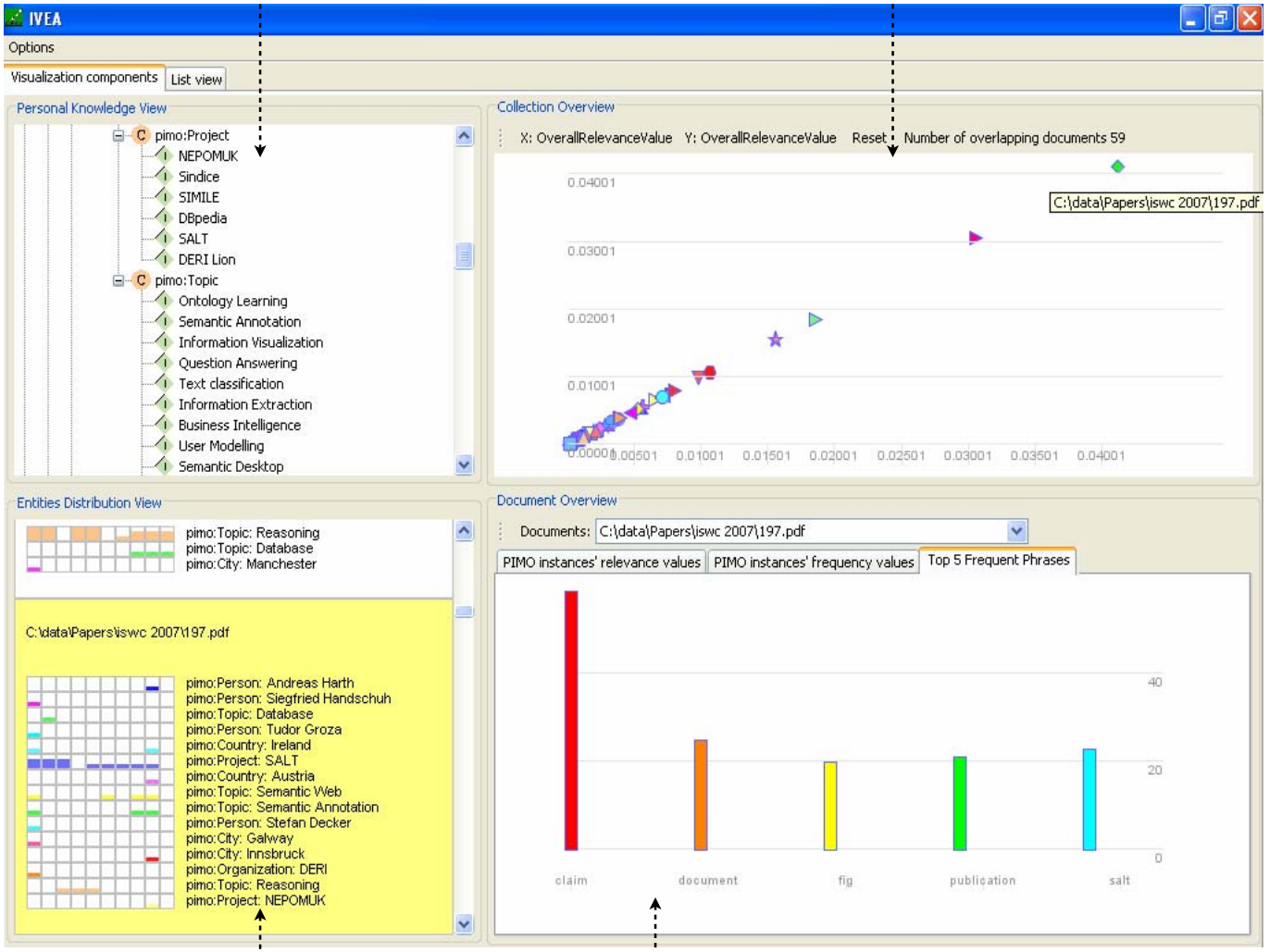
IVEA's visual interface consists of four views as follows:

- **Personal Knowledge View**: The tree-based structure, as shown on the upper-left corner of Figure 4.2, is used to display the concepts, instances and their hierarchical relationships within the user-defined ontology. For each instance, users can defined a list of linguistic variations associated with it. The concepts and instances within this ontology can serve as an anchor for the exploration process. This view also acts as the basis for other interactions within IVEA such as ontology enrichment or modification of instances' linguistic variations.

- **Collection Overview:** The scatter plot, as shown on the upper-right corner of Figure

The Personal Knowledge View shows concepts and instances within the personal ontology

The Collection Overview shows documents containing any of the ontology's instances as glyphs on a scatterplot. Their coordinates are the relevance values of those documents with respect to the ontology's instances.



The Entities Distribution View shows the frequency distribution of ontology's instances within the document being focused on in the Collection Overview.

The Document Overview shows the details of the document being focused on in the Collection Overview. This view has three tabs, each has a bar chart showing information such as top frequent phrases in the selected tab above.

Figure 4.2: IVEA initial interface.

4.2, is used as the overall view to display documents matching the users' interests. On the scatter plot, each document is represented by a dot and its file name is shown in the dot's tooltip text. The coordinate of a dot on the scatter plot is determined by the relevance values of the dot's corresponding document with respect to the classes or instances set on the x and y axes. The initial display of the scatter plot, as shown in Figure 4.2, uses the overall relevance values of documents in the collection with respect to the set of all ontology instances on both axes. Based on this initial display, the users can, for example, see that in the collection being explored, 59 documents overlap with their interests and that the document "*C:\data\Papers\iswc 2007\197.pdf*", represented by the rightmost dot on the scatter plot, is most relevant, based on its coordinate. More details about that particular document can be obtained immediately from the coordinated document overview and entities distribution view as described shortly. Moreover, the dimension of either of the two axes can be changed to reflect how relevant the documents are with respect to the concepts or instances placed on it. The dots' colors and shapes are used to differentiate the associated documents. The dot's size can be customized to accommodate for text collections of different sizes.

- **Document Overview:** Bar charts are used to display detailed views on various characteristics of each document in the collection. Three different tabs containing a bar chart each are used on the lower-right corner of Figure 4.2. The first tab contains a bar chart that shows ontology instances appearing in a document together with the relevance values of that document with respect to them. The second tab has a bar chart that displays the matching ontology instances based on their frequencies. The third tab contains a bar chart that shows the top frequent phrases or terms in a document. For instance, in Figure 4.2, the tab containing the top frequent phrases bar chart of the document "*197.pdf*" mentioned above shows that the word "*claim*" appears most frequently.

- **Entities Distribution View:** This view, as shown in the lower-left corner of Figure 4.2, is based on the TileBars paradigm, originally reported in [56], and some of its variants in [104]. It is used to display the matching ontology instances within each

fragment of a document. The rows are associated with instances whose labels are placed next to them on the right. The number of columns in this view is the number of fragments to divide a document into, whose value can be set by users. Each document is split into sentences and they are put into fragments such that each fragment contains an approximately equal number of sentences. The height of the colored area in a cell is determined by the frequency of the corresponding instance in that particular fragment. By using this view, users can quickly be informed of the relative locations, in which certain entities of interest appear, together with their respective frequencies. For instance, Figure 4.2 shows that the instance "*SALT*" of the concept "*pimo:Project*" appears more often in the first three fragments than in any other parts of the document "*197.pdf*".

4.1.2 Interactions, Manipulations and Coordinations

In line with the visual information-seeking mantra by Shneiderman [120], IVEA provides users with the freedom to interact and control the visual displays in the following manners:

- **Filtering:** It is essential that users are able to filter out uninteresting documents to focus only on a restricted subset of a text collection. In IVEA, users can directly manipulate the overall view displayed in the scatter plot by dragging concepts or instances from the personal knowledge view and dropping them onto the labels of the x and y axes to indicate the dimensions upon which the relevance values of documents are to be measured. IVEA instantly updates the scatter plot by executing dynamic queries for the relevance values of all documents with respect to the instance or the aggregated relevance values with respect to the concept placed on the axes. In Figure 4.3, the x-axis highlights documents relevant to "*Semantic Desktop*" while the y-axis highlights documents relevant to "*pimo:Person*". The relevance value of a document with respect to "*pimo:Person*" is dynamically measured as the aggregated relevance value of that document with respect to all instances of the concept "*pimo:Person*" in the ontology. Hence, the scatter plot can show, among others, documents referring to both the topic "*Semantic Desktop*" and one or more persons who are of specific interest to

users (documents plotted above both axes). The example also illustrates that IVEA can take advantage of the hierarchical relationships between entities within an ontology to allow for the rapid exploration of a text collection at different levels of detail.

- **Details-on-demand:** Once the overall view has been restricted to a specific set of documents, users can investigate the details of each document within that set. Clicking on a dot in the scatter plot will open up the corresponding document in its associated application. Furthermore, to interactively link the collection overview with the detailed views, coordinations between the scatter plot and the bar charts as well as the Entities Distribution View are provided in such a way that hovering the mouse over a dot in the scatter plot will display the corresponding bar charts and highlight the Entities Distribution View of the document represented by that dot. This interaction is also illustrated in Figure 4.3.

- **Personal ontology enrichment:** An innovative feature of IVEA is its capability to enable users, with minimal efforts, to enrich their personal ontologies while exploring a text collection. While investigating a particular document to see how it overlaps with their spheres of interests, users are presented with the most frequent phrases[1] in that document. The top frequent phrases and their corresponding frequencies are displayed in a histogram as shown in Figure 4.4. If any of the presented phrases is of specific interest to users, they are just two-click away from adding it to the personal ontology simply by dragging its respective column in the histogram and dropping on a concept in the ontology. Users have the option to add the selected phrase as a subclass or instance of the targeted concept. Figure 4.4 and 4.5 illustrate how users can enrich the PIMO ontology by adding "*conference*" as a subclass of "*pimo:Event*". We believe that this is of benefit to users as it allows them to update their personal ontologies with new entities interactively. Consequently, they can better explore a text collection when their spheres of interests are better represented. Besides, an extended personal ontology is useful not only for IVEA but might also be useful for other personal ontology-based applications.

[1] Candidate phrases are nouns or noun chunks consisting of at most 3 words.

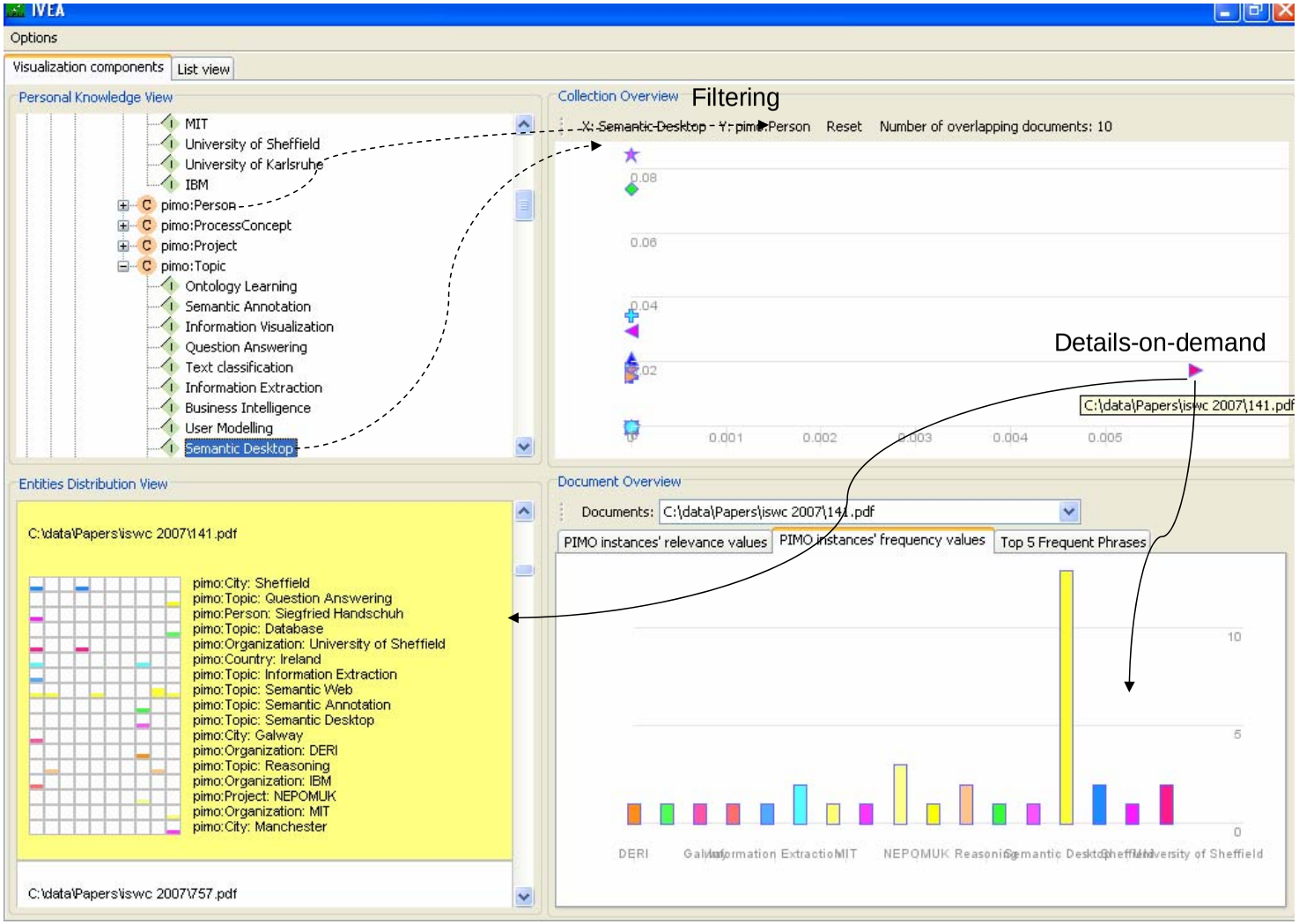


Figure 4.3: Interactive Filtering by the instance "*Semantic Desktop*" on the X axis and the class "*pimo:Person*" on the Y axis, and Details-on-demand by clicking on the glyph representing the document "C:\data\Papers\iswc 2007\141.pdf", resulting in the corresponding values being displayed in bar charts within the three tabs of the Document Overview panel, and that document's Entities Distribution View also gets updated accordingly.

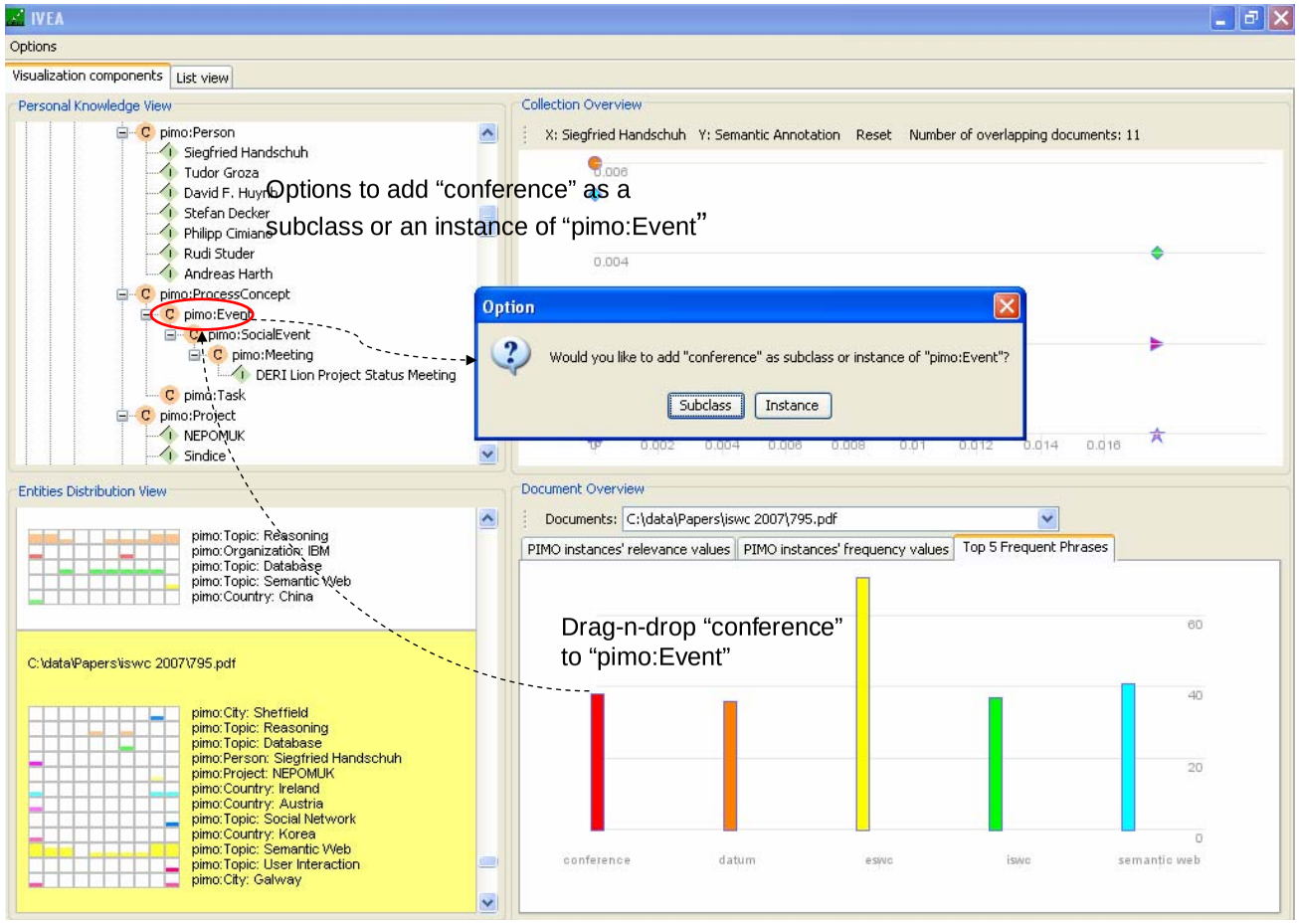


Figure 4.4: Adding new concepts or instances to the PIMO ontology.

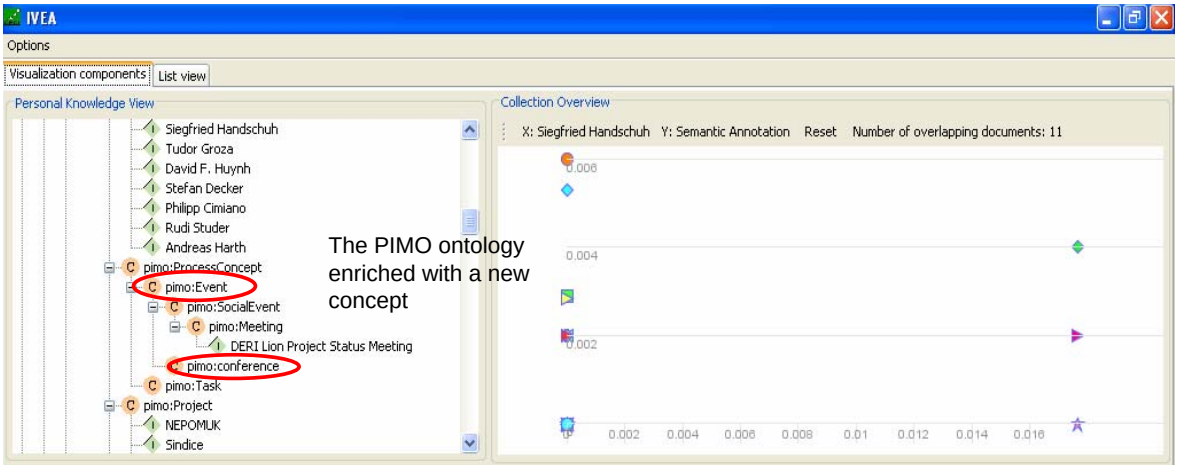


Figure 4.5: The PIMO ontology added with a new concept.

4.1.3 Implementation

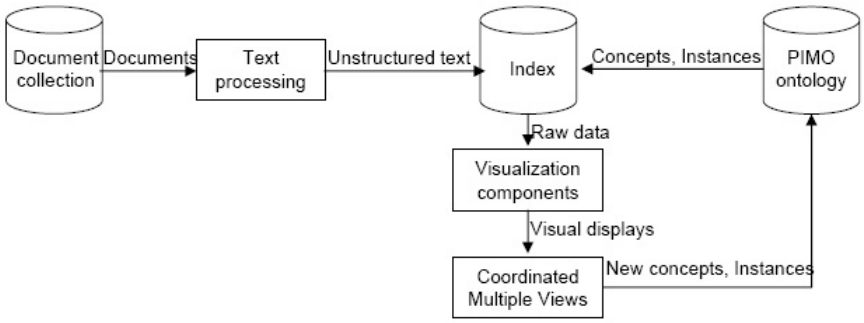


Figure 4.6: Implementation Overview

The implementation overview is shown in Figure 4.6. Documents in a collection are first analyzed by the text processing component, which is based on various natural language processing resources (Tokenizer, Sentence Splitter, POS Tagger, JAPE Transducer) provided by the GATE framework [26]. Each document is split into sentences to identify the fragments' boundaries for use in the Entities Distribution View. Furthermore, to suggest users with entities potentially matching their interest, most frequent noun chunks in each document together with their frequencies are extracted.

The analyzed text is then stored into a Lucene[2] index. A Boolean query (with the default OR operator) consisting of all ontology instances is used to retrieve documents that are of

[2] `http://lucene.apache.org/`

interest to users. The term weight in a document of an instance is used as the relative relevance value of that document with respect to that instance. Here a variant of the well-known TF.IDF term weight function is used, which takes into account the frequency of a term locally (in a document) and globally (in the whole collection). The relative relevance value of a document with respect to a class is the aggregated relevance value of that document with respect to all of its direct instances and recursively, all of its subclasses.

The above relevance values and the frequencies of ontology instances in fragments of each document are used as raw data for the visualization components. The implementation of the visual displays is based on the prefuse library [61] and Java Swing components.

4.2 Formative Evaluation

We carried out a small-scale formative evaluation in which six researchers participated. They performed a number of pre-defined tasks on a list-based baseline interface and then on IVEA. The subjects all had prior knowledge about ontologies and Boolean query operators. They were assumed to have the same sphere of interest, whose concepts and structures were encoded into a pre-defined ontology. This ontology acted as the basis for exploring and analyzing a test collection consisting of 62 research papers in the Semantic Web domain. Although the collection's size was not large, it was suitable for this initial study since it could be representative of some real-world situation, for example, when a researcher wishes to understand different characteristics of a document collection consisting of scientific papers published at a conference in a particular year.

We do not go further into detail of the study here as we realized that it had a number of limitations. A better designed study would (1) involve a larger number of participants who are representative of the target users population (knowledge workers) and hence some may have limited knowledge about ontologies and Boolean query operators, (2) require a larger test collection, (3) use a more extensive set of tasks, and (4) require that each interface be used first by half of the subjects. Nevertheless, at an early stage of IVEA's design and development, this formative study served as an opportunity for some users to interact with the prototype and give us their feedback on IVEA's potential usages as well as usability

problems.

The useful feedback came mainly from the following questionnaire. The participants were asked to give subjective ratings for all questions below, except for the last one which is open-ended. The ratings were on a Likert scale from -2 (very bad/ completely disagree) to 2 (very good / completely agree) with 0 being neutral.

- How well suited are the visualization components to the exploration and analysis of the document corpus?
- How helpful are the visualization components in the exploration and analysis of the document corpus?
- How helpful are the visualization components for enriching the PIMO ontology while exploring and analyzing a document corpus?
- How easy are the visualization components to use?
- How self-descriptive are the visualization components to work with?
- With proper documentation, how well do you think you could use the visualization components in the future?
- How easy to learn do you think the system is?
- Do you consider the personalized exploratory visualization tool as having added value compared to existing desktop search engine?
- Do you think the integration of personal knowledge via the PIMO ontology offers more advantages to the exploration and analysis task than other data browsing tools?
- Do you perceive the suggested top frequent phrases for enriching the PIMO ontology relevant?
- How appealing do you find the design and layout of the visual components?
- Do you perceive the visual components' design and layout as not too complex for the task?

- Are there any functionalities or visual components that you expect to be included with the tool but were not available?

The mean scores of the responses on the questionnaire are shown on Figure 4.7. Overall, the initial user feedback on IVEA was encouraging to further research in this direction. Among the various aspects surveyed, a particularly well-liked attribute of IVEA was that its design and layout were appealing. Furthermore, the participants considered IVEA as having added value to desktop search engines that are readily available on their desktop. IVEA was also perceived to be both helpful and easy to learn. Other aspects that received mean rating scores of 1 (on a scale from -2 to 2) included IVEA being suitable to the task of exploring a text collection, being easy to use, and the personal ontology enrichment feature being considered helpful for the task as well. Meanwhile, the ratings were a bit low for its self-descriptiveness and being complex for the task. Nevertheless, the participants agreed that it could be easily used in the future with proper documentation.

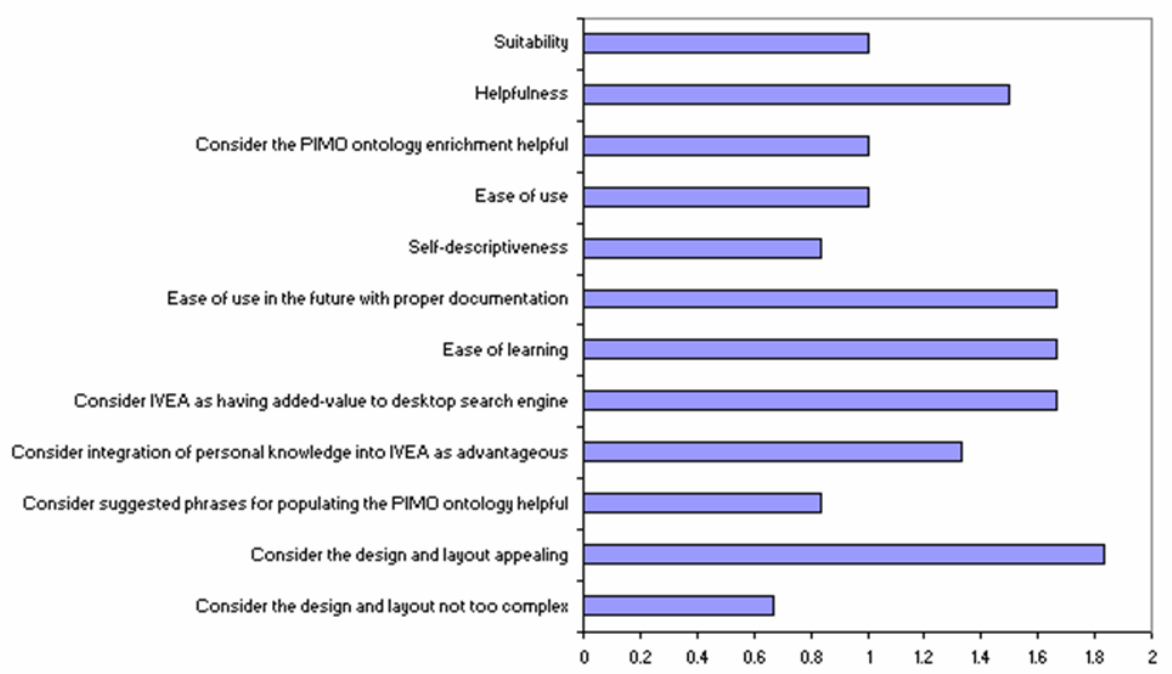


Figure 4.7: Questionnaire responses

The responses for the open-ended question gave us other useful suggestions to improve user experience with IVEA. For the filtering interaction, a participant emphasized the need to allow users to "*multi-select [entities] in PIMO [the user-defined ontology] to the scatterplot*

[for filtering], ...but the scatterplot is 2D, so need something multi-dimensional" (selection of multiple entities at the same time for filtering in a multi-dimensional representation). Although the scatter plot currently employed in IVEA for filtering allowed users to focus on documents which were relevant to the concepts/instances placed on its two axes, it had an inherent limitation in that users could only have an overview of how relevant the documents in a collection were with respect to only *two* entities of interest at any one time. User feedback indicated that the ability to explore a text collection along multiple conceptual dimensions at the same time would be beneficial.

The feedback in the open-ended question also highlighted some shortcomings of the TileBars-based Entities Distribution View. When this view was used to explore a document containing many entities of interest to users, its presentation became harder for the subjects to get a quick grasp of the distributions of many entities. Therefore, the participants suggested some ways to mitigate this problem, such as "*highlight select items in TileBars [Entities Distribution View]*" and "*TileBars – instance list [to be] group[ed] by class, [which would make it] easy to read*". In addition, in this initial version of IVEA, the full text of a document could be shown in a separate display, and hence a mouse-over would result in a tooltip containing the text in the selected fragment, so some participants stated that "*in the TileBars tooltip I would have liked the occurrences of the instances to be highlighted*".

Other useful suggestions included a "*search box for PIMO concepts*" and "*Bar chart – use [the] same color for the same class in [the] ontology or group by class somehow*".

4.3 Summary

In this chapter, we introduce IVEA, an innovative information visualization tool which supports the task of exploratory document collection analysis based on users' interests and knowledge. IVEA leverages upon a personal ontology as a formal representation of a user's interests and knowledge, and it employs multiple coordinated views so that text collections can be explored from different angles. The integration of a personal ontology into IVEA allows users to rapidly explore text collections at various levels of detail. To cater for users' evolving spheres of interests, IVEA enables them to enrich their personal ontologies with

new concepts and instances interactively. A small-scale usability study was carried out and the results were sufficiently encouraging to further research in this direction. The study provided us with useful suggestive indications and valuable feedback to subsequently improve the design of IVEA.

In Chapter 5, we address the limitation on the filtering interaction with an innovative multi-dimensional visualization. In Chapter 6, we will discuss our effort in tackling the visual complexity of TileBars-based Entities Distribution Views so that it is easier for users to quickly understand the distributions of entities of interest within a document. In Chapter 7, we describe how we can provide users with more detailed information than just the distribution of terms, by showing the gist of a term's usage contexts within documents in a visual form that can facilitate comparisons of those contexts between documents.

Chapter 5

Visual Abstraction and Ordering in Faceted Browsing of Text Collections

In Chapter 4, we have introduced the initial prototype of IVEA, an information visualization tool aiming at supporting users in exploring a text collection in such a way that their interests and knowledge can be taken into account. Among its features, IVEA facilitates users in filtering tasks in order to leave out uninteresting documents and focus only on documents containing certain entities of interest. In the initial version of IVEA, this filtering interaction can be performed on a scatter plot, as shown in Figure 4.3. Even though a scatter plot is an intuitive visualization for filtering tasks, it suffers from two main limitations: (1) when used on a large collection of text documents, there might be too many glyphs representing documents to display, hence the occlusion problem arises due to overplotting, (2) users can only explore how documents are related to at most two entities of interest at a time. Feedback from a formative user study on IVEA has suggested that the capability to explore a text collection along multiple dimensions at a time would be beneficial.

In this chapter, we focus on addressing the above limitations in the filtering aspect of IVEA. It is worth noting that as the hierarchy of entities of interest to users, encoded in a user-defined ontology, can be used to characterize documents within a text collection, they can be treated as facets and hence the filtering interaction can be referred to as faceted filtering or faceted browsing. Interested readers can refer to Section 3.3 where we discuss the background of and existing tools for faceted browsing.

In order to design a visualization that can support faceted browsing without the two limitations inherent with scatter plots, next we analyze user requirements for this filtering interaction on text collections. Based on these requirements we will present an early solution with certain limitations, and then go into detail the visual abstraction and ordering techniques that can turn this matrix-based visualization to a suitable solution. We then describe a user study to gauge how this visualization can help users in faceted browsing tasks.

5.1 Faceted Browsing of Text Collections

We were interested in user experience with existing faceted user interfaces, therefore we invited nine persons who were familiar with faceted browsing to participate in contextual interviews. They were asked to demonstrate their recent use of a faceted user interface to achieve a real task of their own. They explained their interactions while we were observing. Afterwards, each of them discussed their experience with using these websites (including amazon.co.uk, yelp.com, landsend.com, cnet.com, daft.ie, dabs.ie). Their feedback can be summarized as follows:

- The facets in these websites were highly appreciated as they were relevant and helpful to narrow down the search space. Some subjects were not happy with the default set of facets, as they did not reflect their most relevant criteria.

- Comparison of items across different facets was very important in many cases for making decisions. To avoid missing the best matches, users had to look at different combinations of facet values inherent in the focus items. Sorting by one facet at a time was therefore not considered effective.

- Facets were not equally important. Some facets were more important than others.

- Having to go through a long list of focus items was time-consuming. This is in line with results from a study on faceted user interfaces [80], which showed that while users spent equally much time on the query, the facets and the results on the first results page, they focused entirely on the items on the second and third results pages.

This suggested that after choosing the facet values to narrow down their options, users still needed to spend a considerable amount of time looking for specific items.

It is worth noting that most of the websites used by the participants only allowed for the single-valued selection mode i.e. only one facet value could be selected at a time, e.g., a brand or a price range (hence conjunctive queries used at each filtering step), only some websites, such as yelp.com and amazon.co.uk, also allowed for multi-valued selection mode, e.g., on the "*Categories*" facet in yelp.com (and hence disjunctive queries used). In the former case, users needed to narrow down the search space in a stepwise manner by choosing only one value in a facet at a time. As such, if none of the items in the results set met their needs, they had to backtrack and explore different paths in order to compare items using a combination of facets that were of import to them to find the best matching items. While comparisons were important to users in making choices, not all sites using the faceted navigation paradigm offered this feature. In the latter case, the multi-valued selection mode catered for vague criteria from users (e.g., multiple neighborhoods were considered acceptable for the location facet while a user was looking for a restaurant), and thus allowed for displaying a larger set of items matching one or more values of a facet. In this situation, users needed to look into the details of each item to figure out which values of a facet an item matched. In both cases, the users' feedback indicated that choosing an item matching their needs involved comparing items based on a combination of facets rather than just one.

While none of the websites used by the participants dealed with documents, they shared a similar feature with most current faceted user interfaces for browsing a text collection, such as the one shown in Figure 3.14, in terms of displaying the results set. Filtering interactions resulted in document items returned in a list and users still had to traverse through many result pages to select a particular document for further analysis. This similarity exists because even though documents are content-bearing items, their contents are completely ignored and only the (usually hand-crafted) metadata is utilized as with other kinds of resources. This has two setbacks:

- It is not always the case that metadata are available for a text collection to be explored, since the effort required to generate the clean metadata (e.g., to classify if a document

belongs to one or more categories) is tremendous on large datasets. An example of this issue is highlighted in [101].

- Even when metadata are available, as in the case of carefully curated digital library collections, the existing paradigm, however, falls short of being adequate to support users in the task of filtering for potentially relevant documents while exploring a text collection. Document contents are only taken into account in free text search to further narrow down the results set. This issue is also highlighted in [20] after the authors conducted an experimental study on a visualization to support search interfaces and realized that users of digital libraries are most often interested in the content of the documents rather than their metadata. The finding suggests that while metadata is certainly useful for filtering and aggregation to obtain basic analytical statistics and relationships, when documents are used, their contents are far richer in terms of useful information that can be explored and hence should not be ignored.

Therefore, we believe that it would be useful to further support users, who are interested in a certain set of entities of interest, to explore a text collection whose metadata are not available. In this case, it is important to:

- *Show how relevant a document is to these entities*. This is due to the fact that unlike the binary relationships often seen between an item and a facet value (e.g., an electronic item and a brand), documents are content-bearing items and hence the relevance relationship is of importance to users in choosing which documents to focus on.

- *Use disjunctive queries instead of conjunctive queries*. When focus items are documents, users should be able to consider multiple facet values simultaneously since a document is usually relevant to a number of concepts at the same time. In addition, while exploring documents, the fact that some documents do not mention certain entities may be of interest to users (e.g., in the Enron email dataset, it may be interesting to identify the set of documents mentioning both "*leak*" and "*investigation*", and the set of documents only mentioning "*leak*" and nothing else in the legal activity category). This feature is lacking in most faceted user interfaces since only a single facet value

can be used at a time and thus facet values can only be ANDed together. Disjunctive queries can return a set of items that do not contain certain facet values and this could be of interest to users. This may be suitable not only to documents but also to other digital resources when *slice-and-dice* operations are required [128], e.g., filtering for photos that satisfy the condition "*Australia and in 2004 and not portrait*" from a photo collection.

- *Show basic repository statistics*. This is to support comparisons between different groups of documents meeting certain criteria. In the example above, it may be useful to compare the sizes of those two sets of emails. The use of disjunctive queries enables comparisons directly on the user interface and also reduces the number of different steps that need to be carried out to achieve this goal.

The above requirements, together with what we can learn from the user feedback on existing faceted user interfaces, beg the question if properties of documents can be visually displayed in such a way that users can make a better informed decision faster than having to traverse all pages of results and looking at one document after another. The key limitation of existing works is the lack of an aggregated view showing how each document relates to each of the entities. While certain issues were raised about faceted user interfaces in general [58], they tend to focus on the display of a large number of facets, e.g., which facets to show (adaptively) when there are many. The issues identified here regarding the focus item representation for faceted browsing of a text collection have largely been left untouched. We argue that user experience can be improved if the following design desiderata are met:

- Each focus item should have a compact representation expressing its correspondences to facet values.
- Users should be able to perform cross-comparisons of focus items over different values of one or more facets.
- The display can be visually abstracted to deal with a large amount of focus items.
- Users should be able to interactively reorder facets based on their preferences, resulting in different displays of focus items.

5.2 A Matrix-based Visualization for Faceted Filtering of Text Collections

To meet the above requirements, we first proposed using a multi-dimensional visualization, one dimension for each facet value, as an alternative to the linear result listing paradigm. This visualization, as shown in Figure 5.1, was inspired by TableLens [103] and FOCUS [124]. Within the matrix, rows represent entities selected for filtering, columns represent documents containing at least one of those entities, and each cell shows the relevance value (a TF-IDF based score) of a document with respect to an entity via its height. To decide a cell's height, we use k-means clustering to identify three clusters of relevance values, and the maximal values of the three clusters are used as thresholds. In addition, there is a cross-hair highlighter which follows the mouse movement to indicate the row and column at the mouse's current position. The vertical part of the cross-hair highlighter helps to focus on which entities a document contains and its horizontal part helps to show the distribution of an entity in a collection. In this visualization, entity de-selection is made as easy as entity selection, i.e. users can de-select facet values by right-clicking on an entity and the whole respective row is removed from the visualization.

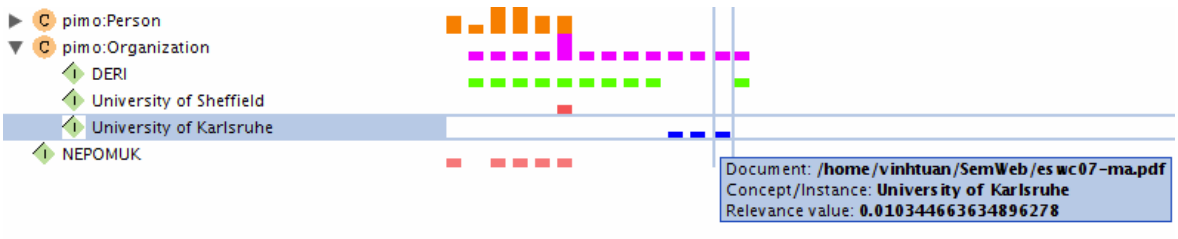


Figure 5.1: The initial matrix visualization for faceted filtering. Here a column corresponds to a document and the height of a cell indicates that relevance of a document to an entity.

This representation, in effect, blends visualization with faceted browsing in that users can select entities from hierarchical facets and then documents relevant to one or more of the selected entities are displayed on the visualization. The relationships between documents and entities of interest to users are automatically obtained based on text analysis of the documents' contents. Since hierarchical relationships between entities are taken into account,

selecting a class will result in the automatic inclusion of all of its direct instances and recursively, all of its subclasses. Thus, facet selection for filtering can be done at different levels of granularity and multiple facet values can be selected in a single operation. With this visualization, each document has a visual representation indicating its correspondences to facet values and users can perform cross-comparisons of documents over each facet value.

Although it meets two of the design desiderata mentioned in Section 5.1 to support faceted filtering interactions, the visualization proposed above provides no visual abstraction to cater for a large number of documents and no interactive ordering of facet values to enable users to easily compare items across different facet values (entities). In the following sections, we focus on the proposed solution for those two issues. Here, documents are information items and concepts/entities of interest to users are facet values.

5.3 Visual Abstraction of Documents

Given the typical screen resolution, too much data is confusing and limiting information manipulation. Therefore, it is important to collapse a visual display when desired so that users can focus more effectively. In this respect, the semantic zooming and document grouping features can be employed simultaneously to achieve a highly flexible visual abstraction when the number of documents makes it challenging for users to digest and explore. We can represent relationships between documents and a set of concepts in a bipartite graph $G = (D, C, E)$ whose vertices belong to two disjoint sets D representing documents and C representing concepts. If $d_i \in D$ is relevant to $c_j \in C$, then there is an edge $(d_i, c_j) \in E$ connecting them, whose weight is the relevance of d_i with respect to c_j.

5.3.1 Semantic Zooming

Hierarchy is an important organizational paradigm, which helps users abstract key concepts from groups of similar items and structure their reasoning [20]. The semantic zooming feature is based on the hierarchical relationships among entities (facet values) and hence helps avoid cluttered interfaces by providing different levels of abstraction. The hierarchy attached with the matrix allows users to dynamically drill down or roll up to achieve views

at different conceptual levels of detail to focus on particular subgroups. This, in effect, is the aggregation of vertices in a set of concepts, hence replaces edges that connect documents to instances of a class with edges that connect documents to that class only. For instance, in Figure 5.2 assuming that c_1, c_2, c_3 are instances of a class, they can be grouped into a single concept c_{123} representing that class. Documents containing any or all of these three concepts is considered as containing c_{123}. Thus, the resulting bigraph on the right has fewer edges. In general, the levels of semantic zooming correspond to the number of hierarchical relationships between entities defined in the ontology. An example is shown in Figure 5.3.

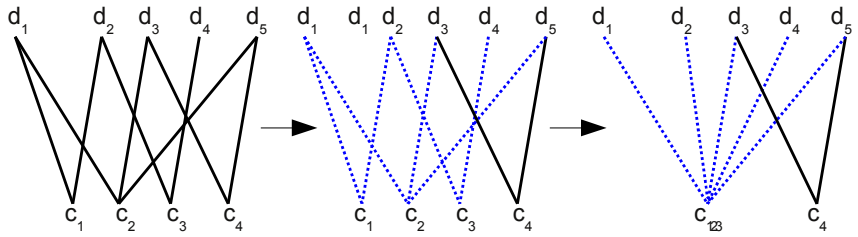


Figure 5.2: Semantic Zooming Bigraphs

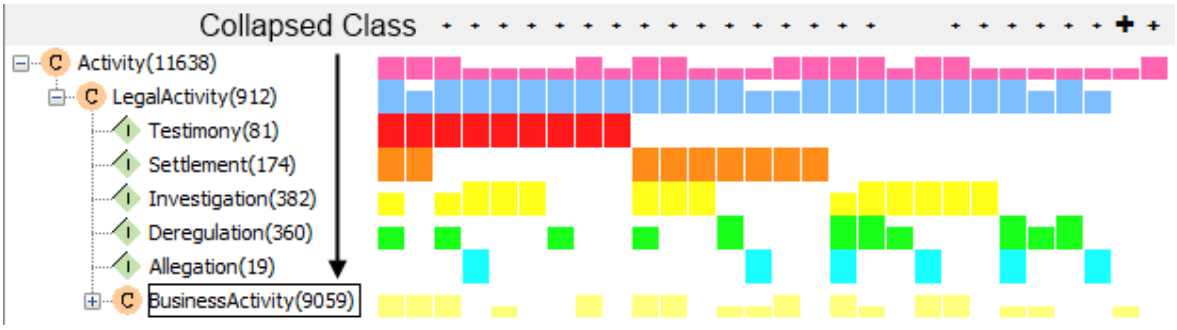


Figure 5.3: Semantic Zooming Example

5.3.2 Documents Grouping

While semantic zooming can abstract away a lot of details, the number of documents in a relatively large collection can still be too much to be effectively displayed on a limited screen space. Here the documents grouping feature provides further abstraction based on the notion of Structural Equivalence of individuals in social networks [84], defined as below.

Definition Objects a, b of a category C are structurally equivalent if a relates to every object x of C in exactly the same way as b does [84].

This notion is used to partition objects in a set into classes of structurally equivalent objects, which leads to the ability to derive a reduced set of categories in which belonging objects are considered equivalent. When these objects are individuals in a social network, the set of derived categories represents "maximal relationally homogeneous groups" [84].

Here, we treat documents and concepts as objects and adapt the Structural Equivalence notion as per below.

Definition Given a set of concepts $C = \{c_1, .., c_n\}$, the set of structurally equivalent documents with respect to C consists of documents containing all elements of C.

In other words, documents d_i and d_j in D are structurally equivalent with respect to C if there exist edges (d_i, c_k) and $(d_j, c_k) \in E$, for $k = (1, .., n)$. Given a set of selected entities C , we can identify a set of structurally equivalent documents with respect to each set in the powerset of C (excluding the empty set). For example, in Figure 5.4 the documents d_1, d_3, d_5 contain c_2, c_4 and not any other concepts. Therefore, they can be put together into a group d_{135} as shown in the right bigraph.

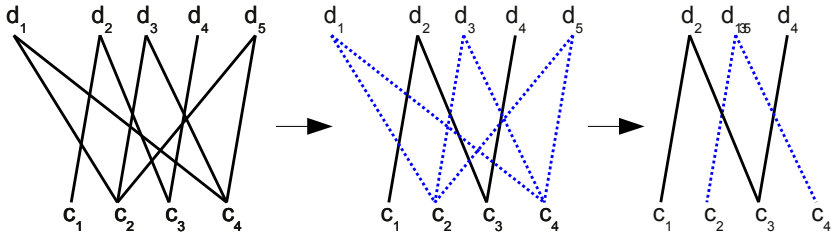


Figure 5.4: Document Grouping Bigraphs

It is worth noting that while collapsing and expanding facets with semantic zooming are already widely employed in many facet browsing websites, the advantage of our approach is that documents grouping can be used **simultaneously** with semantic zooming to achieve visual abstraction on two different dimensions of concepts and documents at the same time, as can be seen on Figure 5.3. As such, this approach enables even more flexibility with regard to the levels of granularity at which information is viewed and manipulated. As shown in

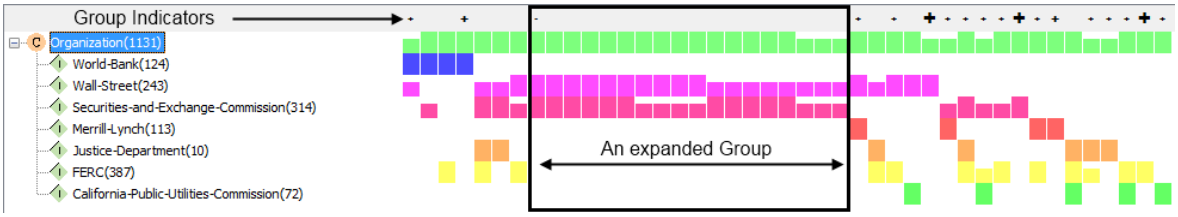


Figure 5.5: Document Grouping Example

Figure 5.5, although the screen space is limited, the visualization can still cope with a large set of documents. Here, the faceted filtering process results in 1131 relevant documents containing at least one of the instances of the concept "*Organization*". However, we do not need to display 1131 columns in the initial matrix visualization. Since 18 structurally equivalent groups are identified, together with 8 documents that contain unique combinations of the entities used for filtering, only 26 columns need to be shown, as only one (randomly chosen) document of a group is initially displayed on the matrix, while other documents that do not belong to any groups are still displayed as a regular column each. In fact, if there are n selected facet values, only a maximum of $2^n - 1$ columns are needed for the initial display. The column of the representative document of a group has a '+' sign on top, which is a visual cue to indicate that there are more documents containing exactly the same set of entities. We also use k-means clustering on different group sizes to find three clusters of size and use the maximal values of the three clusters as thresholds. Thus, the size of a '+' sign can indicate the relative size of a group. Hovering the mouse over a '+' sign will pop-up the exact number of documents in that group. This visual cue overcomes the need to show numeric values, which require varying spaces and can distort the consistent layout of the matrix. Clicking on this representative column will make visible all documents in a group and its visual cue changes to '-' as shown in Figure 5.5. Clicking again on the representative column will hide other documents in that group. This interaction simplifies the comprehension of the visual display of a large number of documents without users having to examine a matrix containing a large number of columns, since the initial display does not depend on the actual number of focus items (documents).

5.4 Visual, Interactive Ordering based on Facet Values

As previously mentioned, not all facets are equally important. For each user and/or for each task, there is an order of importance of facet values accordingly. Therefore, we consider facet values that are placed on top to be more important than those below them. Thus, documents and groups of documents are reordered based on their correspondences with facet values. As in Figure 5.3, in the facet "*LegalActivity*", the value "*Testimony*" has the highest position, therefore documents and groups of documents that are relevant to "*Testimony*" are moved to the left and those that do not are moved to the right. Within these two groups, they are subsequently ordered by their relevances to the second value, "*Settlement*". This ordering is done similarly until the last facet value. This ordering can be efficiently achieved using a bitmap of a document or a group of documents, which is constructed by assigning a 1 bit if a document/group of documents is relevant to a facet value, a 0 bit otherwise, and the first facet value corresponds to the highest order bit. For instance, the left-most document in Figure 5.6 corresponds to the value "11111000", highest among the derived bitmap values.

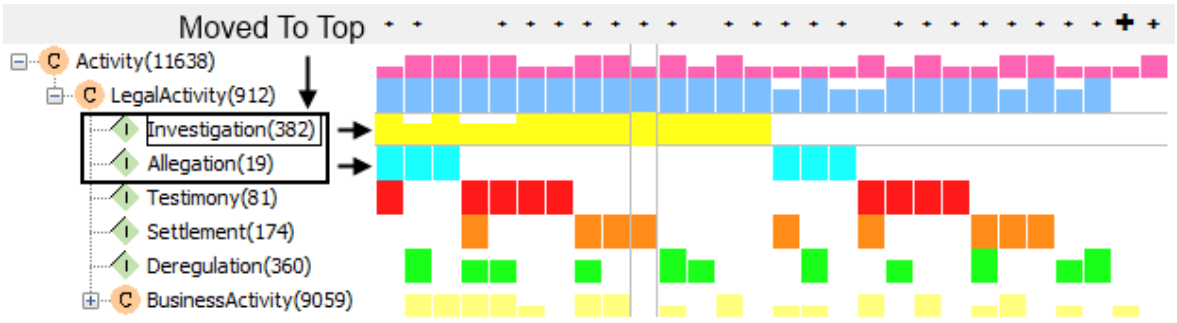


Figure 5.6: Facets Reordering Example. In comparison to Figure 5.3, in this Figure, the two entities "*Investigation*" "*Allegation*" have been moved to the top and the matrix was changed accordingly.

Furthermore, users can interactively reorder facet values while exploring a text collection. As shown in Figure 5.6, in the facet "*LegalActivity*", if the values "*Investigation*" and "*Allegation*" are considered more important, they can be moved (via drag-and-drop) on top. The documents' order within the matrix is changed accordingly as a result (compare with Figure 5.3 to see the difference). We believe that this visual ordering based on facet values enables users to easily compare focus items, in this case documents/groups of documents,

across facet values in a meaningful, prioritized way.

5.5 New IVEA Visual Interface

The updated version of IVEA[1], which incorporates the new faceted filtering feature described in this chapter, is shown in Figure 5.7. In this Figure, IVEA is in use with a collection of more than 38000 emails from the Enron dataset.

A number of new features have been included in IVEA as follows to improve user experience.

- In the Personal Knowledge View (Panel 1 in Figure 5.7), query previews are added to concepts and instances within the ontology to indicate how many documents contain each of those entities. We also incorporated user feedback on the previous version of IVEA (discussed in Section 4.2) to provide a keyword search box on top of the tree structure to facilitate quick navigation to entities when the hierarchy is large.

- In Panel 2, the scatter plot previous used for filtering is now replaced with the matrix visualization described in this chapter. Keyword search is available to further restrict the matrix to display a particular subset of the results set. If the timestamps of documents can be extracted, they can also be used to aid filtering. Filtering interactions can still be performed by dragging entities from the tree structure in Panel 1 and dropping them onto Panel 2. Users can quickly remove all filtering criteria by clicking on the "Clear Filter" button.

- Users can interact with the matrix visualization as described earlier (e.g., expanding or collapsing groups of documents, or reordering entities). Once they wish to focus on a document for more detail, they can right-click on the column representing that document in the matrix. This interaction will result in the Document View (Panel 3) and Entities Distribution View (Panel 4) being updated to display the details of the document in focus.

[1] A screencast is available at `http://resources.smile.deri.ie/projects/ivea/screencast/2010-06/`

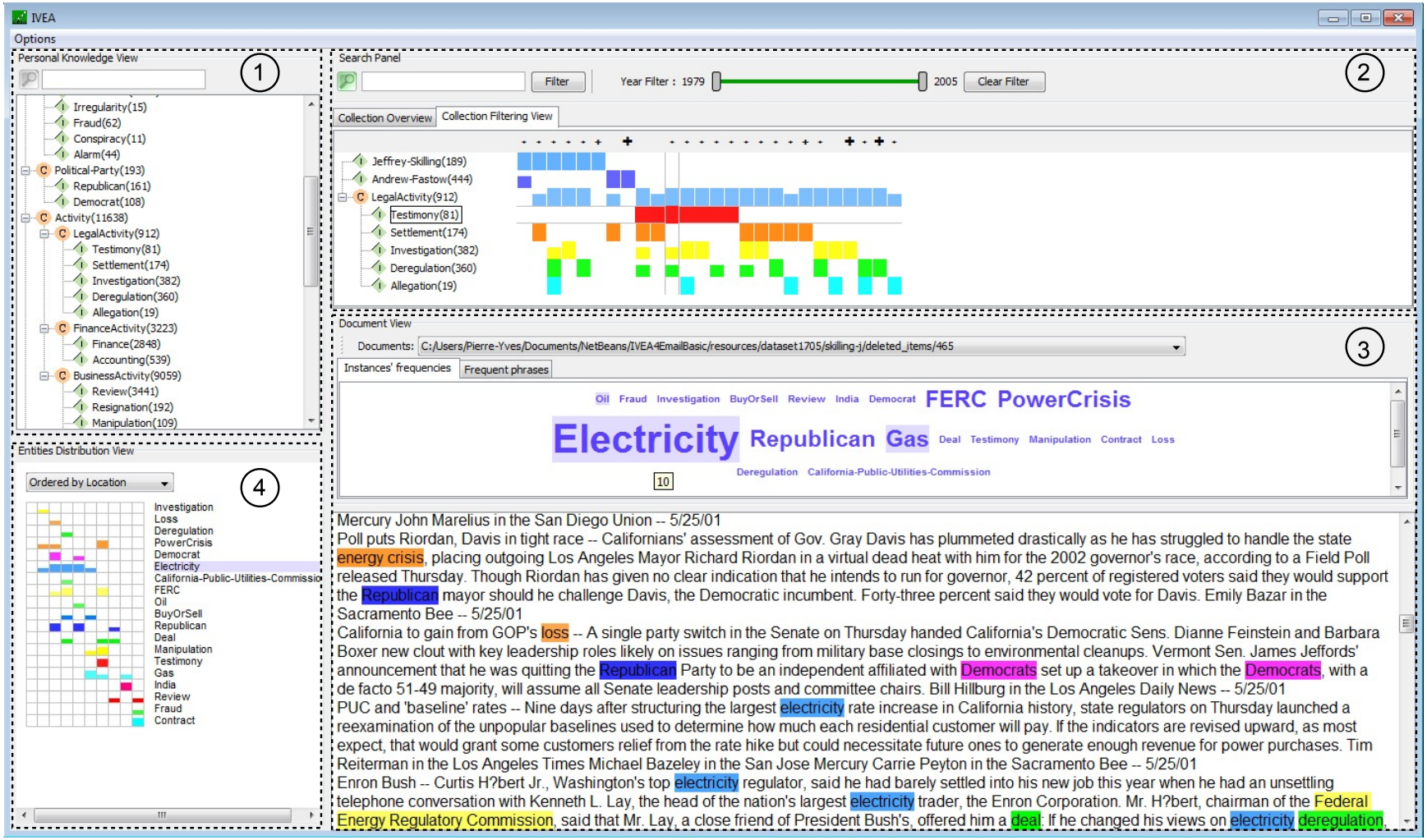


Figure 5.7: The updated interface of the IVEA prototype. In Panel 1, entities of interest to users are shown in a tree structure. Panel 2 is used for faceted filtering. Panel 3 shows the details of a document being in focus among those presented in the filtering view in Panel 2. Panel 4 displays the distributions of entities within the same document in focus as in Panel 3.

- The Document View in Panel 3 has two parts. The upper part of has two tabs: one tab has an entity cloud showing which entities and how many times they appear in this document, the other tab shows a word cloud of the most frequent phrases extracted from the same document. The reason that we use word clouds instead of bar charts as in the initial version of IVEA is that bar charts do not scale well when there are a lot of entities to display. The lower part of the Document View shows the full text of the document in focus, which is also annotated with occurrences of entities of interest. The color palette used for annotations is the same as that used in the matrix visualization in the Filtering and Entities Distribution views to maintain consistency (the same entity gets the same color in all views). The inclusion of a view showing a document's contents helps users navigate to parts of the text containing relevant entities a lot faster than opening a separate view as with the previous version.

- In the Entities Distribution View (Panel 4), previously we showed a list of all Entities Distribution Views for all relevant documents and highlighted the one corresponding to the document in focus. However, this is not scalable for large collections, therefore in the updated version, we only show the view for the document in focus upon selection from the filtering view (Panel 2). Upon a mouse-over on top of an entity in the entity cloud in the Document View (Panel 3), the corresponding row showing the distribution of that entity is highlighted (e.g., "*Electricity*" being highlighted in Panel 4). We will go into more detail on the change we made in the Entities Distribution View in the next chapter.

With this updated interface, we carried out an evaluation on the faceted filtering aspect of IVEA as described next.

5.6 Evaluation

We conducted a user study to test the null hypothesis that *there are no differences on users' performance and preferences on interfaces using the matrix-based representation and interfaces using the linear listing paradigm.*

5.6.1 Method

Participants

18 people (9 females, 9 males) participated in the study, two of whom participated in pilot sessions. Six of them were in the age range of [18,25], nine in [26,40], and three in [41,60]. In terms of occupations, the subjects were accountants, administrative assistants, project administrators, managers, programmers, undergraduate and graduate students in History, Business, Physics, Industrial Engineering and Information Technology/Computer Science. All were familiar with web search and had a need to seek for information from documents (with frequencies ranging from a few times a week to hourly). They had all used commercial websites employing the faceted browsing paradigm. None had any prior experience with the IVEA prototype.

Materials

We used a set of 38180 emails from the Enron email dataset[2] as the test collection in the study. These emails varied in length, from relatively short to long ones. Many of them were internal disseminations of news articles together with further discussion. A set of pre-defined facets and facet values were used for the study. The study was conducted on a 15" laptop with a 1440x900 screen resolution.

Design

The study used a within-subjects design. The independent variables were the interface type and the task type; the dependent variables were the participants' performance and their subjective ratings.

In this study, we implemented four different interfaces to display the results set returned from faceted filtering interactions:

- Simple linear listing (UI1): In this baseline interface, as shown in Figure 5.8, items were displayed linearly, ordered by their relevance to the set of selected entities (facet

[2]Available at `http://www.cs.cmu.edu/~enron/`

values). Each result item had an email subject, a set of entities that it contained among those used for filtering and a short text snippet. This interface resembled current faceted browsing websites in terms of showing which facet values an item is associated with for a multi-valued facet.

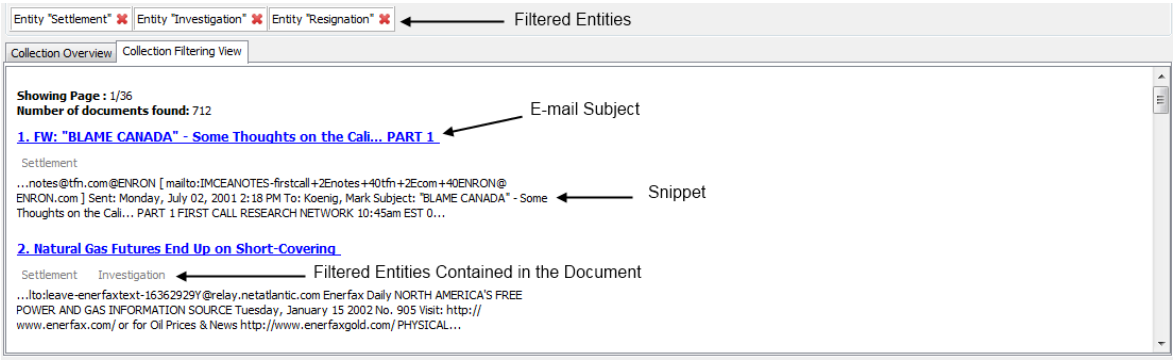


Figure 5.8: UI1 - A linear listing view of the results set.

- Linear listing with grouping (UI2): As shown in Figure 5.9, this interface was similar to UI1 in that it also employed the linear listing paradigm that users were already familiar with. However, instead of showing the names of the entities contained in each result item, small squared icons were used as their visual representations. Entities' names could be known from the legend or the tooltip text. Each square was filled with a color based on its corresponding entity and its relevance value, in the same fashion as cells in the matrix representation. In addition, items were also ordered and grouped in a similar fashion to our proposed approach in Section 5.3 and 5.4 in that items containing exactly the same set of entities were displayed continuously and ordered by their binary signatures. For instance, in Figure 5.9, the first document has binary signature of "111" because it contains all three entities, the second one has a signature of "110" because it only contains the first two entities. Documents having the same signature of "110" will be listed next, and followed by those whose signatures have smaller binary values. This interface served as a strong baseline.

- Matrix-based view (UI3): This interface employed the proposed approach in Section 5.3 and 5.4.

- Matrix-based view with linear listing (UI4): As shown in Figure 5.10, this hybrid

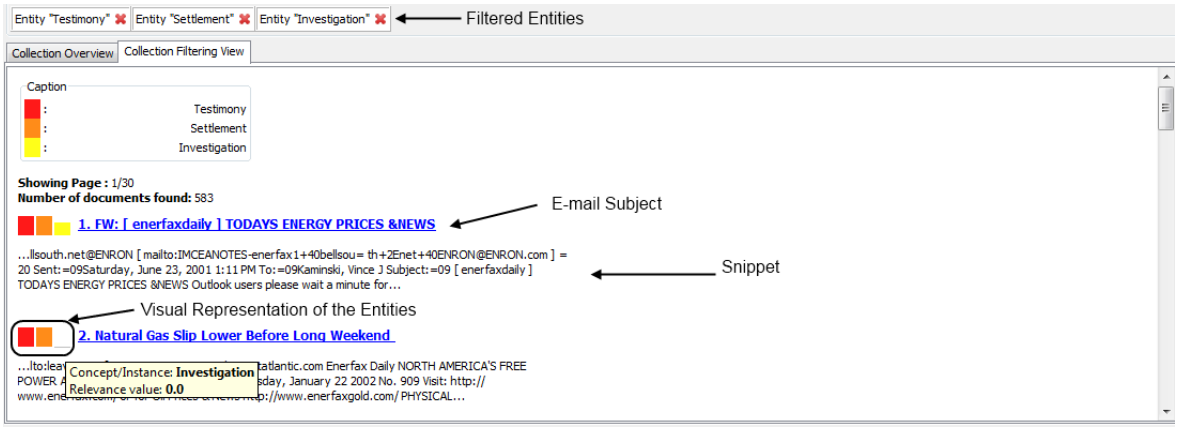


Figure 5.9: UI2 - A linear listing view of the results set, added with visual indicators of which entities are contained in each document.

interface was a combination of UI3 and a slightly-modified version of UI2. When users clicked on the representative column of a group of structurally equivalent items, they were not expanded as columns in the matrix. Instead, this group of documents were shown in a linear listing view in which items were ordered by relevance with respect to the set of entities shown in the representative column. The representative column was visually highlighted in the matrix to indicate which combination of entities were contained in that group of items. The visual indicator for each document was made more compact as non-relevant entities were removed, which resulted in a more focused representation while the correspondence between a document and the whole set of selected entities was still visually available in a larger context in the matrix.

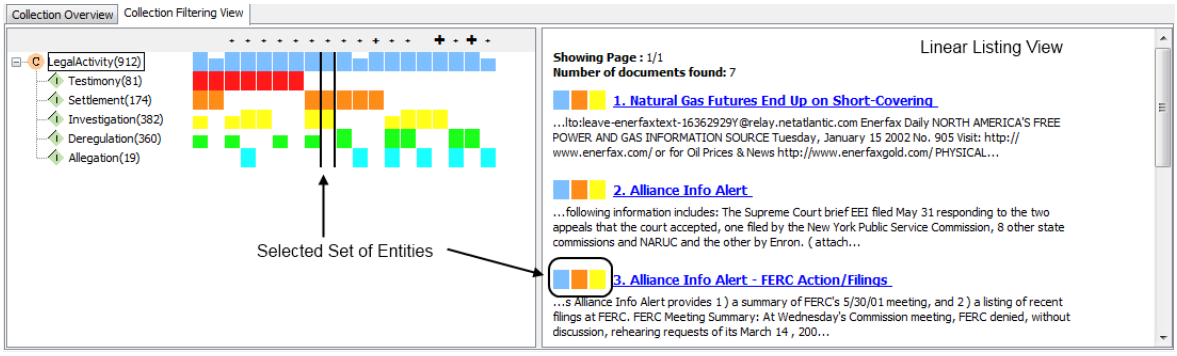


Figure 5.10: UI4 - A hybrid version of the matrix-based visualization and the linear listing view.

The above four interfaces were considered to cover a range of variations from the one that

resembled the current linear listing paradigm, to linear listing enhanced with icon representations, which was also grouped and ordered, to the proposed matrix-based representation and a hybrid version of matrix-based and linear listing view. During the study, each implemented interface was referred to by a neutral, color-based name, e.g., "FacetExplorer-Green" for UI4.

Procedure

The subjects were given a short briefing about the study. They were also introduced to and shown video screencasts of all four different interfaces and asked to carry out a number of sample tasks on each of them. They were not informed of which ones employed our proposed approach. Then they were asked to perform four different sets of tasks, one for each interface. To eliminate learning effects, the order of the interfaces was counterbalanced. The assignments of the four different sets of tasks to the four interfaces were also counterbalanced using Latin Square as suggested in [68].

Each set of tasks consisted of the below three parts:

- The first task was related to basic repository statistics, i.e. asking for the number of documents meeting some criteria in two subtasks.
- The second task asked the subjects to filter for (i.e. identify) specific documents meeting some criteria including a time period.
- The last task was an optional task involving the subjects freely filtering for documents of interest.

The first two tasks were preceded with short excerpts from news articles from the New York Times, which were related to the Enron event and hence the dataset, to set the contexts. The contexts were helpful to provide the subjects with some topically interesting background information and hence they could relate to the situation and be motivated to carry out the tasks [79]. In addition, we designed the first two tasks such that they required the same amount of efforts (the same number of steps) to find out the answers i.e. all subtasks involved the same number of different entities, however their associated Boolean operators

were varied to avoid learning effects. While the first two tasks enabled us to compare users' performances with respect to obtaining basic repository statistics and identifying the relevant set of documents meeting certain criteria, the third task was an opportunity for the subjects to freely experiment with various features of the interfaces being evaluated and hence we could solicit their comments. As such, we took into consideration the task completion time and accuracy in the first two tasks and sought qualitative results from the third one. The participants did not have to write down answers as doing so may introduce large variations in terms of time. They only pointed to them on the screen and verbally informed the facilitator, the accuracy and task completion time were extracted in post-hoc analysis based on the screen and voice capture.

Below is one of the four sets of tasks used in the user study[3]:

Task 1. "*...Even as the celebrations unfolded, accountants and trading experts at the company's Houston headquarters were desperately working to contain a financial disaster, one that threatened – and ultimately would destroy – everything Enron had become. A handful of executives were struggling to sound the alarm, but with Enron's confidence in its destiny, the warnings went unheeded...*"

In this context, you may want to focus on emails that mentioned accounting and Kenneth L. Lay. Among them,

- How many emails mentioned loss ?
- Among those that did not mention loss, how many emails in 1999-2002 mentioned alarm.

Task 2. "*...On Aug. 14, stunning the market, Jeff Skilling announced he was resigning after just six months as chief executive, citing undisclosed personal reasons. He left assuring investors that the finances of Enron had never been better...*"

- You may want to find out more about internal communications surrounding this event. Identify emails in 2000-2004 which mentioned Jeff Skilling, resignation, and the role Chief Executive Officer.

[3]The contexts are excerpts from: `http://www.nytimes.com/2002/02/10/business/enron-s-many-strands-company-unravels-enron-buffed-image-shine-even-it-rotted.html` (Accessed November 2010)

Among them,

- Identify emails that mentioned investigation
- Identify emails that did not mention investigation.

Task 3. Filter for emails of interest based on either of the two contexts above.

In order to compare the subjects' performance when the Latin Square design was used, it was important to take the tasks' characteristics into consideration while comparing them. While the number of entities involved were equal across tasks, the sizes of the results sets were not, and that was something we could not control. To answer the structured tasks, the subjects had to process results sets of different sizes and as larger results sets may require more time, we should not attribute that to the interfaces themselves. In addition, the accuracy of the answers should also be considered, since participants who made mistakes should be considered as having lower performance scores. Therefore, we employed a metric similar to the *qualified search speed* proposed in [68], called *relevant browse speed* (RBS)[4], which reflected the number of focus items the subjects were able to process per minute, weighted by the accuracy of their answers.

$$\text{RBS} = \frac{\text{Results Set Size}}{\text{Task Completion Time}} \times \frac{\text{2*Recall * Precision}}{\text{(Recall + Precision)}} \tag{5.1}$$

Once the subjects finished the tasks on all interfaces, they were asked to give their subjective ratings on each of the interfaces, based upon the following adjectives: "*Easy to use*", "*Easy to browse*", "*Stimulating*", "*Satisfying*", "*Overwhelming*", "*Flexible*", "*Self-descriptive*", "*Organized*","*Tedious*". The ratings were on a Likert scale from 1 (Strongly disagree) to 9 (Strongly agree), with 5 being neutral. Similar to the study conducted in [159], we used a wide range to have a more sensitive testing instrument. In addition, the reason the subjects were asked to give these ratings once they finished all tasks was to give them a chance to interact with all interfaces, and therefore they could calibrate their ratings in an informed manner. Previous work showed that the order of interfaces could have an effect on the

[4]In [138], we defined RBS using recall values in the second part of the equation, here we used F1 values to take into account both recall and precision values. The differences were negligible and the findings presented in the next section were the same for both ways of defining RBS.

subjective ratings (e.g., in [159]) and hence we want to avoid this effect. Upon completion of tasks on all four interfaces, the subjects were asked to answer an overall questionnaire:

- Do you think the tasks are equally difficult? (apart from the different results set sizes).
- Which interface is the easiest to use?
- Which interface helps you learn the most about the document collection while exploring?
- Did the matrix-based presentation of result items change the way you used to explore a collection of text documents?
- Can you describe an example where the matrix-based presentation of result items helped / hindered, frustrated your exploration process?
- Do you think the grouping of documents containing exactly the same set of entities helped or hindered exploring a large number of relevant documents? (both groupings in list views and in matrix views)
- Do you think the ability to abstract the view on the matrix simultaneously on two dimensions (concepts and documents) helped or hindered the document exploring process?
- Do you think that the interactive ordering in a matrix-based presentation of result items helped or hindered while you are trying to have a clearer idea of the results set? Can you describe an example ?
- Would it make a difference if the matrix only used one color ?
- What was your experience when browsing/ filtering freely without any guided tasks in order to gather documents potentially containing relevant pieces of information?
- Overall, how would you rank the interfaces? Please provide comments on your ranking.

The findings are presented next.

5.6.2 Results and Discussion

Before we discuss the findings, it is worth pointing out a few challenges we faced in setting up the tasks for this experimental study. As faceted navigation is usually used to support exploratory activities, constructing tasks for evaluation is known as a challenging issue [79]. Among the desired characteristics proposed in [79], exploratory tasks should indicate ambiguity in information needs, but also suggest a knowledge acquisition, comparison or discovery task. As such, a combination of structured and unstructured tasks are usually used in experimental studies on faceted browsing (e.g., [159]), with "known item" search tasks also used in some cases (e.g., [80]). While it may be better to evaluate a tool supporting exploratory search in a longitudinal study, our evaluation effort here was not on the tool (IVEA) itself as a whole, but specifically on its way of presenting items from the filtering step. Therefore, we designed the filtering/comparison tasks in such a way that they could be carried out on all four interfaces. Although the prototype could provide more features than reflected in the structured tasks, these features were emphasized in the introduction and the unstructured tasks instead. In a controlled user study, the type of pre-formulated query tasks such as those used here was considered acceptable for comparing different ways of presenting the results set, since they helped to control the variations potentially caused by many elements of an experimental study and to focus on a particular interface component instead of the whole tool [68, 59]. In this discussion, for clarity we refer to the interfaces using their abbreviations, e.g., UI1 instead of the neutral names used in the study.

Quantitative results

When the subjects carried out the tasks, six of them gave up with UI1 on all tasks as they found it too tedious to continue. The rest of the participants tried but also gave up on at least one task with UI1. Therefore, we did not have sufficient data on the subjects' performance on UI1 to include it in further analysis.

The mean relevant browse speeds by UI and task are shown in Figure 5.11. We analyzed this dependent variable using a 3x2 (3 interfaces x 2 task types) repeated-measures ANOVA. The results showed that:

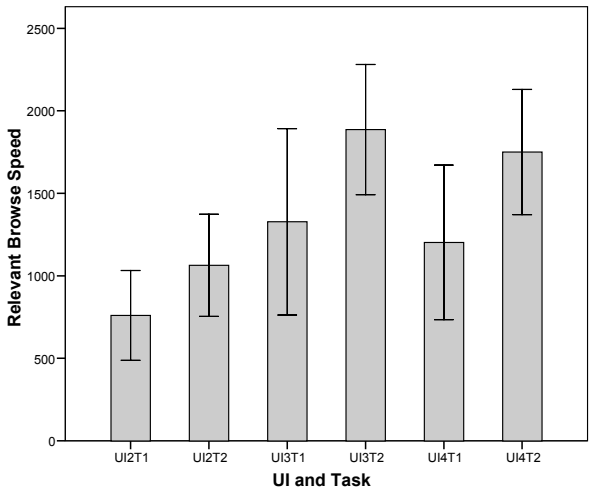


Figure 5.11: Mean relevant browse speed by UI and task.

- The **main effect** of **interface type** was **significant**, $F(2,30) = 4.559, p < .05$, partial $\eta^2 = .23$, hence the null hypothesis was rejected. Pairwise comparisons with Bonferroni adjustment showed that the relevant browse speed on UI2 ($M = 912, SD = 129$) was significantly less than that on UI3 ($M = 1607, SD = 191.8$) and UI4 ($M = 1476, SD = 170.1$), both $p's < .05$. There was no significant difference between the relevant browse speeds on UI3 and UI4. The reason that the mean relevant browse speeds seemed to be high was because groupings (in all three interfaces) and ordering (in UI3 and UI4) allowed the subjects to skip many of the irrelevant groups of focus items. Thus, it was not the case that the participants looked into every single focus item to decide if it matched or not, but only at the visual encodings of representative items of the groups. As such, even when there were thousands of documents to be processed, only a small number of them needed to be considered instead. This reflected the advantage of using grouping and ordering.

- The **main effect** of **task type** was also **significant**, $F(1,15) = 18.347, p < .05$, partial $\eta^2 = .55$. Pairwise comparisons with Bonferroni adjustment showed that the relevant browse speed on task 1 was significantly less than that on task 2, $p < .05$. This may be attributed to the fact that in each task set, task 1 was always before task 2 on all interfaces, so the subjects became more familiar with the interfaces as they carried out the tasks and faster on task 2 as a result. In addition, in task 2, there were always two

subtasks that used the same set of entities but different Boolean operators attached (for the purpose of comparison), hence the participants did not have to form new queries (and hence needed not process a new results set), they just needed to find relevant answers from the same set of results.

- The **interaction effect** was **not significant**, $p > .05$.

Furthermore, the mean values of the participants' ratings on the degree to which they agreed with the adjectives used to describe the four interfaces are shown in Figure 5.12. In general, UI3 and UI4 received more positive ratings than UI1 and UI2 in almost all measures. Below we discuss the ratings for each adjective based on ANOVA analysis and post-hoc tests:

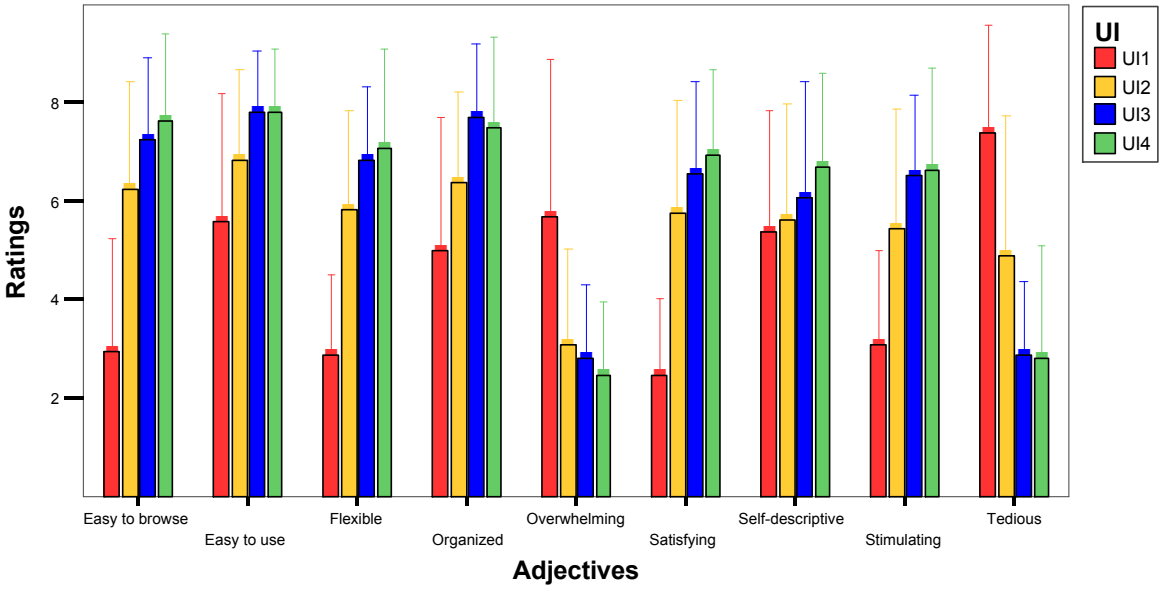


Figure 5.12: Mean subjective ratings.

- "*Easy to browse*": The effect of interface type was significant, $F(3,60) = 18.29$, $p < .001$. Tukey HSD post-hoc tests showed that it was significantly less easy to browse on UI1 than on the other three interfaces, $p < .001$. (No other pairwise differences were significant.)

- "*Easy to use*": The effect of interface type was significant, $F(3,60) = 5.44$, $p < .05$. While the significance value for homogeneity of variances in this case had $p < .05$,

the Welch and Brown-Forsythe tests were both significant with $p < .05$, therefore we could reject the null hypothesis that there was no difference in the mean ratings with respect to the "*Easy to use*" aspect. Games-Howell post-hoc tests showed that UI1 was significantly less easy to use than UI3 and UI4, $p < .05$.

- "*Flexible*": The effect of interface type was significant, $F(3,60) = 18.12$, $p < .001$. Tukey HSD post-hoc tests showed that UI1 was significantly less flexible than the other three interfaces, $p < .001$.

- "*Organized*": The effect of interface type was significant, $F(3,60) = 6$, $p < .05$. Games-Howell post-hoc tests showed that UI1 was significantly less organized than UI3 and UI4, $p < .05$.

- "*Overwhelming*": The effect of interface type was significant, $F(3,60) = 7.66$, $p < .001$. Games-Howell post-hoc tests showed that UI1 was significantly **more** overwhelming than the other three interfaces, $p < .05$.

- "*Satisfying*": The effect of interface type was significant, $F(3,60) = 18.87$, $p < .001$. Tukey HSD post-hoc tests showed that UI1 was significantly less satisfying than the other three interfaces, $p < .001$.

- "*Self-descriptive*": The effect of interface type was **not** significant, $p > .05$.

- "*Stimulating*": The effect of interface type was significant, $F(3,60) = 10.44$, $p < .001$. Tukey HSD post-hoc tests showed that UI1 was significantly less stimulating to use than the other three interfaces, $p < .05$.

- "*Tedious*": The effect of interface type was significant, $F(3,60) = 14.45$, $p < .001$. Games-Howell post-hoc tests showed that UI1 was significantly **more** tedious than the other three interfaces, $p < .05$.

In summary, it was faster and more accurate for the subjects to filter/explore with the interfaces that employed the matrix-based representation (UI3 and UI4) than the linear listing paradigm enhanced with visual encodings and grouping (UI2). The simple baseline, linear

listing only interface (UI1) was perceived to be too tedious to be useful for this kind of activity (which was also well reflected in the subjective ratings). Both the interfaces that used the matrix-based representation were well-liked by the subjects, as they were considered to be easy to use, to browse, flexible, organized, satisfying and stimulating. While the strong baseline interface (UI2) received lower ratings on these aspects on average, the differences were not statistically significant in comparison with UI3 and UI4. The pure baseline (UI1) was also considered to be overwhelming. There was no statistically significant difference in terms of self-descriptiveness for all interfaces.

Qualitative results

The subjects' answers to the overall questionnaire on the four different interfaces are described below. The finer-grained, elaborated details from the participants were useful as more could be understood about their preferences.

The participants' answers indicated that they thought the tasks were equally difficult. When asked which interface was the easiest to use, nine subjects chose UI4, one chose UI3, four of them thought UI3 and UI4 were equally so and only two chose UI2. Furthermore, six subjects chose UI4 as helped them learn the most about the collection, two chose UI3, five of them thought UI3 and UI4 equally did, one chose UI2, while one chose UI2/UI4 because of the listing, and one did not know. All participants thought the use of the matrix-based representation changed the way they used to explore a collection, with positive and encouraging comments such as "*I think that the ability to look at the nodes [document-entity relations] is really really good. This is often the way I wanna look at things but it's not easy to do [with existing interfaces].*" The participants' feedback for questions on interfaces' features are summarized in Table 5.1 via an example for each type of responses, ordered by the non-exclusive Count figures indicating how many subjects mentioned similar comments.

The subjects also commented on their experience when browsing/ filtering freely without any guided tasks in order to gather documents potentially containing relevant pieces of information. Most were really positive in the case they were enjoying exploring the collection, such as "*Easy enough to get the information out to get an overview, especially with UI3 and UI4*" or "*It was fine, but it's hard to start. It allows for a lot of trials and errors, adding*

Description	Example	Count
The document grouping feature helped.	"*It's faster, it's far easier, you can see pretty much all documents, which ones are relevant and which ones are not.*"	15
The facet reordering feature helped.	"*If you were looking for something that is in the document and something that is not in the document, you just put the one that you want at the top.*"	13
Scrolling through lists is tedious.	"*UI2 type of grouping is not as helpful [as Matrix grouping in UI3 and UI4] because you have to scroll down the list. I would give up after a while.*"	8
The matrix provides clear Boolean statements.	"*It made it much easier to answer the type of queries you ask. Mentions X, Y but not Z for instance. And also I think it doesn't require a lot of experiences in logic queries. You just see what's there and what's not there.*"	5
The matrix speeds up search.	"*Yes, I think it helped me to fasten my way of finding something, especially with a huge collection of documents, so I wouldn't mind having something like this on Google website.*"	4

Table 5.1: Comments on Interfaces' Features

and removing concepts, when you try something and then you get rid of it, the overall result will be more satisfying." on UI4. Another subject was surprised about her findings when trying to find documents mentioning a combination of some entities in particular: "*There were certain categories where I thought there would be overlap but there wasn't.*" Some responses were more reserved when the commenters did not know much about the collection (the Enron event) e.g., "*The documents were not interesting for me, if it was something that I would need, then I would probably very much enjoy it.*", and did not define the set of entities of interest themselves. As a result, not all participants performed the unstructured tasks on every interface. In addition, we observed that:

- Asking the subjects to carry out unstructured tasks in a controlled experimental study was troublesome, as the nature of the unstructured tasks were contradictory to that of the settings. The subjects were not in a state of mind that would normally encourage them to freely explore a text collection.

- With our evaluation focusing only on the visual representation to support exploring the results set returned from the faceted filtering step, it was not easy for the subjects to

Rank	UI1	UI2	UI3	UI4
1		12.5%	6.3%	**81.3%**
2		6.3%	**75%**	18.8%
3		**81.3%**	18.8%	
4	**100%**			

Table 5.2: Overall Rankings by Percentage of Participants.

Description	Example	Count
Disliked UI1	"*With UI1 it's nearly impossible when you have a lot of results.*"	16
Preferred Matrix views over List views	"*Overall, it's incredibly powerful [...] I've never seen that kind of visualization before, and it's really good.*"	9
Preferred UI4 over UI3	"*Compared to UI3, the additional list view is really helpful.* "	9
Preferred UI3 over UI4	"*You don't need to click on the column to go to the documents.*"	2
Preferred UI2 over Matrix	"*Because the way it is ordered make the user think more about what she wants.*"	2
Preferred List views over Matrix views	"*It gives you directly the list of items.*"	1

Table 5.3: Comments on Interface Comparison

resist using the prototype as a whole.

Furthermore, the overall rankings by percentage of participants of the four interfaces are shown in Table 5.2. It is clear that UI4 was most liked, followed by UI3. UI1 was disliked by all participants. The rankings were in fact surprising to us, as prior to the user study, we believed that UI3 had the advantage of not requiring users to process more information. But as it turned out, the additional list view in UI4 was perceived as providing more useful information. Further feedback on their justification for the rankings is summarized in Table 5.3.

In terms of other usability aspects, the subjects were also asked if they think it would make a difference if the matrix only used one color. The responses were unanimous, e.g. "*Yes, it would be less effective*", "*It would be more difficult without colors*". Finally, certain usability issues were raised by the participants as described in Table 5.4. Since their frequency was low, we believe that with proper documentation/training, the subjects would be able to overcome these issues once they get more familiar with a new visual representation.

Description	Example	Count
The matrix's + sign is unclear	"*I don't really get why you have bigger or smaller + signs. It makes it more difficult to click if it's small. It's a bit tricky.*"	3
Unclear that a column is a group	"*It's easy to forget that a column represent a group of documents. While interacting with the interface I kinda forgot that that was a group of documents.*"	1
The matrix gets confusing	"*It could get confusing for instance if you have 2 white spaces. Or if the relevance is really less, you might mistake it for a white space.*"	1

Table 5.4: Comments on Usability Issues

5.7 Related Work

In Section 3.3, we have covered notable advances in faceted browsing. In this section, we briefly summarize works that are closely related to ours and discuss how our work is different from them.

A number of faceted navigation systems support exploratory tasks by visualizing the number of items matching the concepts (used as facet values) such as RelationBrowser++ (using bar charts) [160], Elastic Lists (using list items' sizes) [130] and VisGets (using tag cloud color intensities) [33]. Meanwhile, in IVEA, we use the query previews [100]. The main difference between the above systems and our work is the results set presentation. All three systems mentioned above employ the linear listing paradigm on the results set once certain concepts have been selected. In this approach, there are no indicators of which concepts an item matches and how relevant it is to each of them.

ContentLandscape [130], FacetLens [81] and DocuBrowse [46] are applications that visualize results set distributions using Treemaps or Treemap-inspired representations. The ContentLandscape application aims at supporting resource analysis by allowing users to see the coverage of the selected resource set (for instance, if all product groups are represented by the results set) as well as to split the results set to up to three dimensions and compare various measures [130]. This application does not target documents as resources and hence uses conjunctive queries as often seen in other applications. FacetLens [81] is also a faceted browser which employs a more text-oriented variation of the Treemaps paradigm to show the results set, augmented by a timeline showing basic statistics. FacetLens shares certain

visual features with its predecessor FacetMap [123], for which certain concern about the usage of screen estate was raised in [58]. DocuBrowse is a recent addition to faceted browsing research, which targets at usage in an enterprise setting [46]. It takes advantage of the document collection file structure (organizational hierarchy) to support search and browsing. In this enterprise context, DocuBrowse also focuses on file types as genres (spreadsheet, presentation slides, etc.) and these genres, among others, can be used as facet values. Both FacetLens and DocuBrowse do not offer the visual abstraction and ordering capabilities like ours.

It is also worth noting Camelis [40], which is a faceted browser that allows flexible query formulations via various navigation modes. This feature is advantageous when different Boolean operators can be associated with facet values, for example, filtering for photos that satisfy the condition "*Australia and not portrait*" from a photo collection. Most faceted browsers do not provide this feature despite it being an important one. However, while Camelis targets at supporting agile browsing of a document collection, it does not treat documents as content-bearing resources and there is also no grouping or specific ordering available for the results set.

In comparison with other faceted browsers for semi-structured data, such as mSpace [114], /facet [62], BrowseRDF [96] (discussed in more detail in Section 3.3), in our work we provide a matrix-based visualization which can deal with large results sets via visual abstraction based on semantic zooming and document grouping, and enable users to prioritize this presentation based on facet values. To the best of our knowledge, these features are not available in the notable faceted browsers for RDF datasets listed here. However, in comparison to these works, conceptually in our work the faceted browsing component works only on the "*Document contains Entity*" relationship. In Section 8.2, we will discuss how we can possibly extend our work to include richer semantic relationships both in faceted browsing interactions and entities distribution analysis.

Finally, the use of a matrix-based representation to show the correspondence between concepts and documents was proposed earlier in the TileBars paradigm to show distribution of query terms within full text documents [56]. The TileBars approach, however, uses a separate visual representation for each document and its purpose is different from ours.

5.8 Summary

In this chapter, we have proposed a novel approach to support users in exploring a text collection by blending a multi-dimensional visualization with the faceted navigation paradigm. Our proposed approach is based on a matrix representation which is analogous to the already familiar spreadsheet paradigm. Its visual simplicity makes it usable as it can convey the correspondence between a document and a set of concepts. The semantic zooming, document grouping and facet reordering features, which can be used simultaneously, enable users to deal with a large amount of documents as well as to provide users with a substantial amount of flexibility in terms of choosing the levels of abstraction and priorities of the chosen facet values. We conducted a user study on a variety of interface types and the results showed that those that used the matrix representation were more capable of supporting users in exploratory tasks and were also well-liked by the participants of the study, with the hybrid interface receiving the most favorable ratings.

Chapter 6

Reordering TileBars-based Entities Distribution Views with the Barycenter Heuristic

In the previous chapter, we proposed an approach toward faceted browsing of text collections to support users in filtering for documents of interest from a large collection. The next suggested step in the Shneiderman's mantra is for users to see the "*details on demand*", i.e. focus on a particular document to gather potentially relevant information.

In this respect, previous research has shown that in order to gain a deep understanding from a document, "*not only the words that are used are important but also the context they are used in, and understanding them in the context is difficult to achieve algorithmically*" [74]. Therefore, when different features of a text document are treated as a whole, it results in "*a smoothing of passages with an unusual trend, camouflaging interesting patterns across the text*" [74].

As part of the IVEA application described in Chapter 4, we proposed an Entities Distribution View, as shown in the panel in the lower left corner of Figure 4.2. This is the resulting view once users narrow down a collection using the filtering feature, and then focus on a single document among those in the results set. This view can provide users with "*details-on-demand*" about the distribution of entities of interest and a quick way to navigate to the corresponding relevant parts. Its design is inspired by the classic TileBars paradigm [56].

In this chapter we report on our effort to improve the visual presentation of TileBars-based Entities Distribution Views. In this approach, a simplified version of the reordering technique based on the barycenter heuristic for edge crossing minimization in bigraphs proposed in [122] is used to re-arrange elements of a TileBars-based Entities Distribution View in order to provide users with better focus and navigation while exploring a document. We also describe a user study which was conducted to gauge users' efficiency in using the reordered Entities Distribution View.

6.1 TileBars-based Entities Distribution Views

A TileBars [56] visualization can be used to show the distribution information of query terms within full-text documents (see Figure 3.19 for an example). Each document has an associated iconic representation which is a large rectangle consisting of smaller squares. Rows are used for query terms and columns indicate their relative positions within the text. Squares represent TextTiles and show frequencies of query terms via various shadings of grey. This compact representation allows users to swiftly interpret the relative length of a document, the frequency and distribution of each term in it [56].

Within IVEA, we employed a variation of TileBars to show the distribution of entities of interest to users within a document (shown here on its own in the left-hand side of Figure 6.1 for an email from the Enron dataset). Our version differs from the original paradigm in its visual codings in the following ways:

- Every document is evenly divided into ten fragments (this number can be configured by users), each with a relatively equal number of sentences. The reason we experiment with this segmentation scheme is that while the original TextTiling approach [55] used in TileBars can convey the relative lengths of documents, it may not work well for very long documents as the large number of fragments can be counter-effective in producing a useful level of visual abstraction. The reordering approach proposed in this chapter is, however, independent of the segmentation scheme used.

- Each entity is assigned, instead of a grey level, a color from a palette and its associated

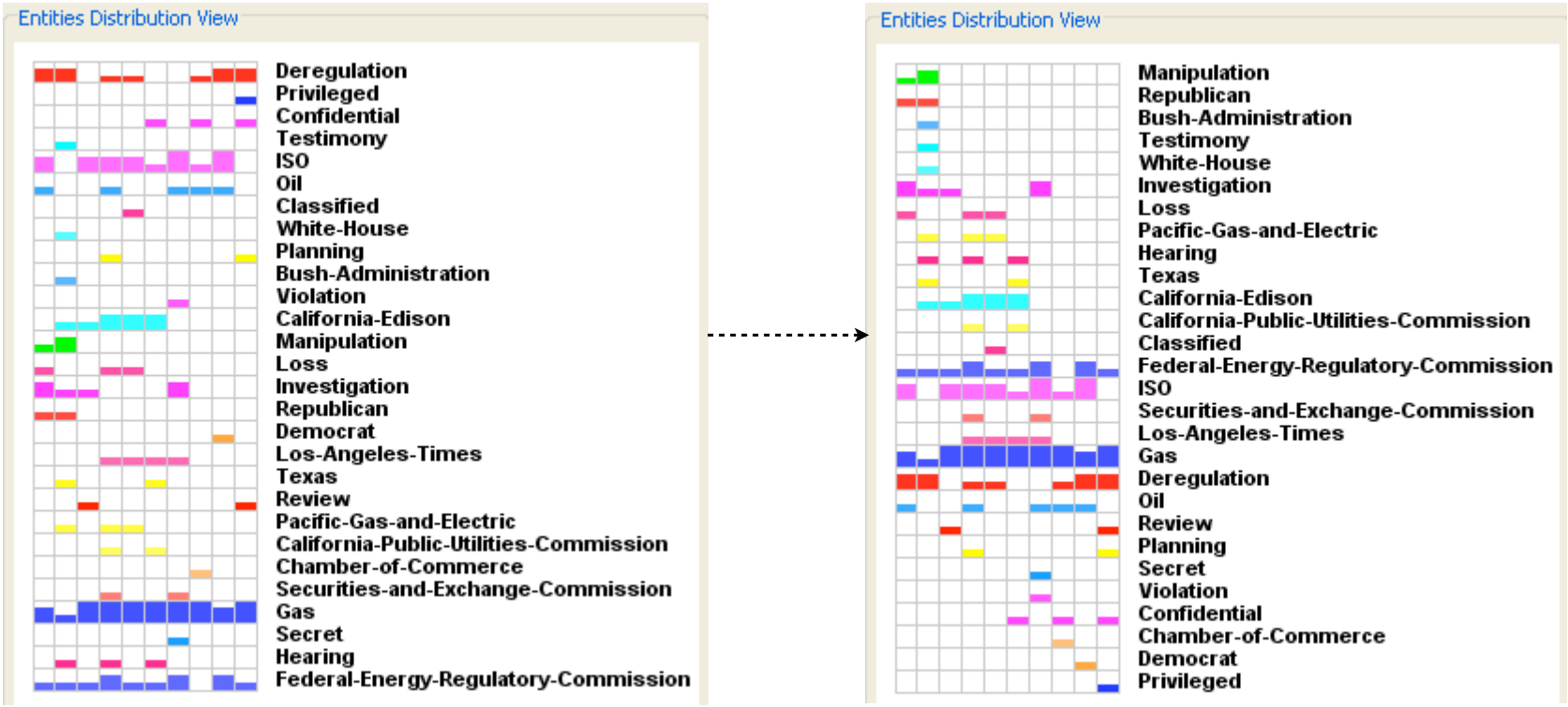


Figure 6.1: The Entities Distribution View on the left displays the distributions of entities in a random order from top to bottom. The barycenter heuristic reordering is applied to this view and results in the reordered view on the right.

row is color coded accordingly. Not only does this help improving visual perceptions, it also helps highlighting appearances of entities in text with their corresponding color in a consistent way. It is worth noting that the original author of TileBars notes that further research shows that "*the aesthetic preference for color outweighs the need for accuracy*" [59].

- An entity's frequency within a segment is indicated by the height of a bar in each square, similar to one of the TileBars variants in [86]. However, instead of using a fixed set of threshold values, we use k-means clustering to identify the maximal values in three clusters of frequency values and use them as thresholds. As such, this approach gives a better discriminative power to the threshold values. In effect, each row is a bar chart of an entity's frequencies in different fragments of a document. Given that the sequential order of fragments is important, we believe that bar charts could deliver better visual effects than arrays of squares with varying degrees of grey.

- While one of the features of the original TileBars paradigm is to compare multiple documents with respect to a set of query terms, here we only employ this paradigm for a details-on-demand view of a single document within multiple coordinated views of the IVEA application. Therefore, comparisons of multiple documents are not our focus in this work.

Similar to the traditional TileBars visualization, in this view users can also click on a square to navigate directly to the corresponding text fragment within a document. This click-to-navigate capability helps users gain quick access to the text portions surrounding entities of interest without having to traverse through a long document.

6.2 Re-ordering Entities Distribution Views with the Barycenter Heuristic

The view shown in the left-hand side of Figure 6.1 is effective in indicating the distribution information of entities across a document, as well as in facilitating quick navigation

based on its horizontal dimension. However, it also shows a shortcoming of the original TileBars paradigm when adapted to show the overlap between a document and users' sphere of interest containing a relatively large number of entities. A TileBars-based Entities Distribution View becomes harder to read when it has many rows, i.e. this happens when users need to use the information on its vertical dimension in order to get an idea of which entities are mentioned relatively in the beginning, the middle, the end or some specific portions of a document. Since there is a diversity in terms of distribution of entities in the text, it is not an easy task to aggregate this information in a quick glance at the view in the left-hand side of Figure 6.1. This poses difficulties to analysts who need to make sense of the Entities Distribution information in a timely manner.

As a TileBars-based Entities Distribution View has a matrix layout in which the order of certain elements can be changed, we propose that its elements be reordered, so that the distribution information can be presented in a more helpful fashion. If we treat the TileBars layout as a connection matrix of a bigraph which consists of (1) a set of entity vertices, (2) a disjoint set of fragments and (3) edges connecting fragments containing entities, the task of reordering a TileBars' elements resembles that of bigraph co-clustering. In bigraph co-clustering, partitioning one set of vertices will induce a specific clustering of the other set of vertices, which then itself induces a new clustering in the first set [31]. This task is combinatorial in nature and hence there are a number of heuristics to partition a bigraph to identify co-clusters, most notably spectral graph partitioning [31] and barycenter heuristic for bigraph edge crossing minimization [133]. The latter has been shown to be more efficient than other methods given its complexity [2]. This approach repeatedly permutes rows and columns of a bigraph's connection matrix such that their barycenters are in an increasing order. As a result of this process, non-zero elements of the connection matrix tend to be placed from the top-left to the bottom-right corners [122]. As such, it has also been considered as a fast and simple way for interactive cluster analysis of reorderable matrices [122].

However, it is important to note that, unlike other reorderable matrices, a TileBars-based Entities Distribution View has a unique constraint in that its rows can be freely reordered while its columns have to appear in a fixed sequence to reflect the linear structure (from the beginning until the end) of text within a document. Thus, instead of reordering rows and

columns of a matrix in turn as usually done in reorderable matrices [122], we propose using a simplified version of the reordering technique based on the barycenter heuristic reported in [122], or one-sided reordering, on TileBars-based Entities Distribution Views. In this operation, only the set of entities are reordered, i.e. moved vertically, while the order of the fragments remains untouched. The goal of this operation is to enable users to clearly see which entities are only concentrated in either the beginning (the top-left corner of the view) or the end (the bottom-right corner) and which entities are distributed across a document (the middle part).

The barycenter of a vector $v = (v_1, .., v_n)$ is defined as [133]:

$$B_v = \sum_{i=1}^{n} i v_i / \sum_{i=1}^{n} v_i$$

Let M_0 be a matrix corresponding to a TileBars-based Entities Distribution View with an initial random order of entities (rows), $K(M)$ be the number of crossings in a matrix M (further details about calculating $K(M)$ can be found in [133]), and $\beta_R(M)$ be a row reordering operation such that their barycenters are ordered from the smallest to the largest. Algorithm 1 shows the modified version for one-sided matrix reordering.

Algorithm 1 One-sided reordering with barycenter heuristic

1: $M^* = M_0$, $K^* = K(M_0)$
2: $M_1 = \beta_R(M_0)$
3: **if** $K^* < K(M_1)$ **then**
4: $\quad M^* = M_1$, $K^* = K(M_1)$
5: **end if**
6: **return** M^*

The resulting reordered view is shown in the right hand-side of Figure 6.1. In comparison with the view of the same document in the left hand-side of Figure 6.1, permuting the rows of a TileBars-based Entities Distribution View suggested an improvement in its visual presentation, whereby filled cells are distributed along the diagonal. This might be of benefit to users, for example, entities which are mentioned through out the course of the text flow of a document may provide insights over the development and evolution of these entities, or the concentration of a particular set of entities co-occurring in some parts of the text may

provide users with useful information.

The studies reported in [122] have suggested that the reordering technique based on the barycenter heuristic is helpful for interactive cluster analysis tasks on reorderable matrices. However, as we mentioned earlier, a TileBars-based Entities Distribution View has a unique constraint in that its columns cannot be reordered, therefore much less can be done in terms of re-arranging elements of this view compared to the reorderable matrices studied in [122]. Therefore, in the next section, we report on an evaluation to see if the simplified reordering operation only on the rows of a TileBars-based Entities Distribution View can provide users with a better focus and navigation support.

6.3 Evaluation

We conducted a user study to validate the hypothesis H1: "*The reordered Entities Distribution Views produce better results in terms of users' performance than the baseline views.*"

6.3.1 Method

Participants

Fourteen people (six females, eight males) participated in the study, two of them participated as pilot subjects to ensure that the given tasks were clear, equivalent and could be completed. Their average age was about 32, ranging from 22 to 41. In terms of occupation, they included an accountant, administrative assistants, project administrators, managers, programmers and postgraduate students. All were familiar with web search and had a need to seek for information from documents (with frequencies ranging from a few times a week to hourly). None had any prior experience with using TileBars.

Materials

We used twelve documents, which were relatively long emails containing news articles inline together with further discussion, from the Enron corpus. The Entities Distribution Views showed the distribution of a set of entities (from a pre-defined ontology) within the

selected documents. Six documents contained from 22 to 27 entities, and the other six contained from 3 to 7 entities. The study was conducted on a 15" laptop with a 1280x800 screen resolution.

Design

The study used a within-subjects design. The two independent variables were (1) the view type, i.e. the baseline Entities Distribution View in which entities were placed randomly into rows versus the view whose rows are reordered by Algorithm 1, and (2) the number of contained entities, i.e. on documents containing a large number of entities versus those containing only a few entities. The dependent variable was users' performance.

Procedure

The subjects were given a short introduction to the baseline view and then asked to perform a set of tasks. The participants were not told prior to performing the tasks that there were two different views being used. After carrying out the tasks, they were shown a reordered view and a baseline view of an example document side-by-side, answered a follow-up user satisfaction questionnaire on the reordered view regarding: suitability, ease of use, self-descriptiveness, easy to learn, confidence, design and layout, and conformity to users' expectations.

On each document, participants performed a task which consisted of 3 subtasks to identify entities that were mentioned (1) only in the first 3 fragments, (2) only in the last 3 fragments and (3) through out the course of the text flow (defined as mentioned in at least 7 different fragments). These subtasks reflected real-world situations in analytical activities, which involved users deciding which parts of a document to navigate to and focus on based on the distribution information of entities of interest within a document. Hence they enabled us to compare users' performance on the two views. It is worth mentioning that other factors such as color-coding and histograms were used in exactly the same way in both views.

Each participant performed the above task on 12 documents, in which 3 documents were used in each of the 4 conditions (e.g., using the reordered view for documents with a large number of entities). To eliminate learning effects, the 12 documents and the 2 views were

introduced in different orders such that their combinations were distributed equally with regard to the two levels of the number of contained entities and the two views.

For each subtask, the F1 score ($2 * precision * recall/(precision + recall)$) was used as a measure of a subject's effectiveness, while a subject's efficiency was defined as the ratio between the effectiveness and the time taken to complete that subtask. A subject's efficiency in a task was calculated as the sum of her efficiencies in all 3 of its subtasks. Efficiency was used as the user performance indicator, which took into consideration both the effectiveness and the task completion time, i.e. it gave higher overall scores to users with high effectiveness scores and a low completion time; but lower overall scores to those with low effectiveness scores and a high completion time. As the subjects worked on three documents in each of the four conditions, the efficiency score for each subject in a condition was the average of the efficiency scores on those three documents.

6.3.2 Results

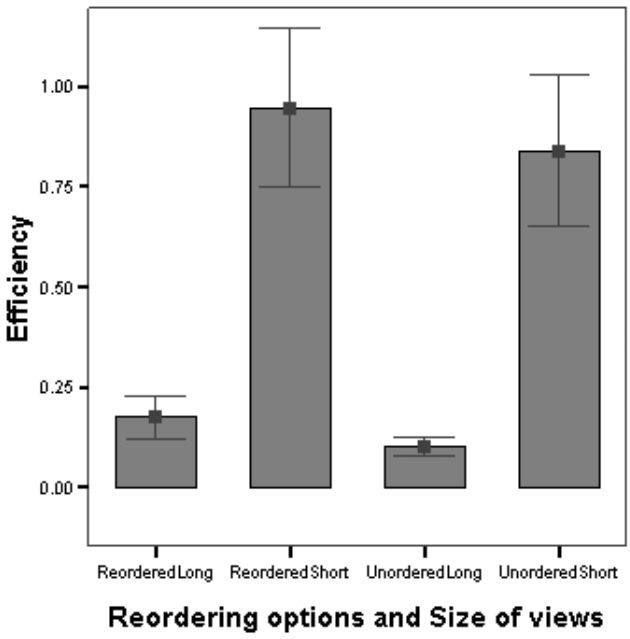


Figure 6.2: The mean values of subjects' efficiency scores in different settings.

The participants' efficiencies, as shown in Figure. 6.2, were subjected to a 2-way within-subjects ANOVA having two levels of view type (baseline, reordered) and two levels of

number of entities (few, many). Note that in Figure 6.2 we use the word "Short" to refer to views with few entities, "Long" to refer to views with many entities, "Reordered" to refer to views upon which the reordering operation was carried out and "Unordered" otherwise.

The results[1] showed that only the main effect of the number of entities was statistically significant, with $F(1,11)= 100.710$, $p < .001$. The main effect of view type was, however, not statistically significant, and yielded an F ratio of $F(1,11)= 3.937$, $p = .073$ (even though $p > .05$, it was very close to the .05 threshold). Therefore, H1 could not be validated as there was not enough evidence to reject the null hypothesis. The interaction effect was not statistically significant, $F(1,11)= .310$, $p > .05$.

The mean values of user satisfaction ratings on the reordered views are shown in Figure 6.3. These ratings are on a Likert scale from -2 (very bad / completely disagree) to 2 (very good / completely agree).

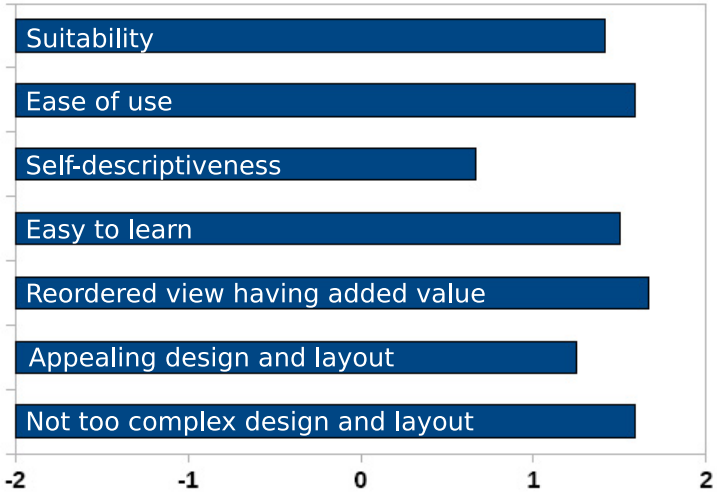


Figure 6.3: User satisfaction ratings

Even though the mean efficiency scores of the subjects when using the reordered views were not statistically significantly higher than those when they used the unordered views, all

[1]The results reported here are different from those that we initially reported in [136]. In [136], we took the efficiency scores of the same subject on three documents in the same condition separately. This, however, means that the samples in each condition were dependent on each other, and in this case ANOVA analysis cannot be performed on the data. Therefore, to address this issue, here we take the average of the efficiency scores on three documents in a condition to represent a subject's performance score in that condition, as suggested at: `http://www.graphpad.com/articles/interpret/principles/population.htm` - last accessed date: 27 April 2012.

of the subjects agreed that the reordered Entities Distribution Views had added value in comparison to the baseline view, and hence the highest mean score over all the user satisfaction ratings as shown in Figure 6.3.

Examples of comments from the participants include:

- "*I can really see the benefits of reordering. It makes it a lot faster to navigate.*"
- "*Search for words with proximity: if they appear close together, it might mean something; if they are apart, may not mean much.*"

Other aspects of the reordered view were also well-liked by the subjects, as they thought that it was suitable to support exploration and analysis of a document's contents, it was easy to use or to learn, and they also considered its design and layout to be appealing and not too complex. Meanwhile, the ratings were lowest on self-descriptiveness. The common feedback was that subjects needed to be given an explanation on what the columns represented. If there were no introduction at all to this kind of view, they would have needed to experiment with it for a while to figure out. All agreed that with a documentation detailing the introduction they got in the study, they would be able to use it comfortably on their own.

Other helpful suggestions that might improve the Entities Distribution Views include: ordering the view by entities' name alphabetically, putting fragment numbers as x-axis, adding report generation capability (i.e. generating a report consisting of all paragraphs / fragments containing an entity), the use of colors might need to be more distinguishing as similar colors might make it confusing, etc.

6.4 Related Work

Our work here is related to many other works in visual analysis of document structures and term distributions discussed in Section 3.4 in Chapter 3, with TileBars being the main inspiration for the proposed visualization. The key difference between our work and others is that we focus on a visualization to be used to provide details on demand, i.e. the distribution of entities within a specific document and not to compare how terms are distributed across different documents.

6.5 Discussion

The TileBars paradigm was originally introduced in the context of showing the distribution information of query terms in documents. As the number of terms in a query is usually small, there are also few rows within a TileBars. Hence, no previous research work has focused on reordering a TileBars' rows to provide a better presentation when the number of rows becomes large, for instance, when used to show the distribution of a large number of entities of interest to users within a document. While TileBars-based visual interfaces are capable of providing users with an overall picture of the distribution of a set of entities all at once, the visual complexity issue makes it difficult for users to make sense of this information. Therefore, in order to improve the user experience, it is important to transform this visual information into a more digestible display.

It is worth noting that reordering rows of a TileBars-based Entities Distribution View is not a trivial task. Despite its constraint of columns having to be in a fixed sequence, there are still $n!$ different permutations to be considered for a view with n rows. Our proposed approach employs a simplified version of a reordering technique based on the barycenter heuristic [122], which is computationally efficient. This subtle yet powerful operation results in a more coherent presentation of the entities distribution information and hence can provide a better support for users in identifying patterns of occurrences of certain entities over the course of a document's text flow. Despite the differences in the quantitative performance scores of the participants in the user study being not statistically significant (which might potentially be due to the small number of participants), the qualitative responses from them suggested that the reordering operation did bring added value compared to the baseline version. Therefore, this visualization would benefit users who have the need to visually analyze and navigate documents in a timely manner.

While our choice of the barycentric ordering technique is due to both its simplicity / efficiency, which is critical in interactive visualization applications and its compelling visual layout, it may be interesting to compare other non-arbitrary orderings (e.g., alphabetic, first occurrences, sum of frequencies) vs. barycentric ordering on TileBars-based Entities Distribution Views. Apart from the reordering operation, the hierarchical relationships between

entities might also be employed to interactively abstract the view for an even more compact representation.

6.6 Summary

The classic TileBars paradigm has been used to show the distribution information of query terms in full-text documents. However, when used to show the distribution of a large number of entities of interest to users within a document, it hinders users' quick comprehension due to the inherent visual complexity problem. In this chapter, we present our effort to improve the visual presentation of TileBars-based Entities Distribution Views, in which a simplified version of a reordering technique based on the barycenter heuristic for bi-graph edge crossing minimization is used to re-arrange their elements. The reordered view enables users to quickly and easily identify which entities appear in the beginning, the end, or throughout a document. A user study has been undertaken and its results suggest that even though the quantitative performance scores were not statistically significantly better with the reordered views, they were well-liked by the study's participants and these participants also appreciated the benefit of the reordering operation while exploring a text document.

Chapter 7

Topic-based Content Abstraction for Visual Concordance Analysis

7.1 Visual Concordance Analysis

In the previous chapter, we have described our effort in providing users with the ability to understand the distribution of entities of interest within a document using a TileBars-based visualization. While this feature is useful for users to quickly navigate to and understand the dispersion of entities within a document, and consequently, any interesting co-occurrences of entities, it is insufficient in supporting users in analyzing the text that is written around a term within a document, or comparing how a term is used in one document versus another, or at one point in time versus another. As the linguist J. R. Firth puts it: "*You shall know a word by the company it keeps*" [41], this kind of analysis / comparison can be achieved by looking into the usage contexts of the term in question by analyzing words immediately surrounding it within the text. In this analysis, a concordance is commonly used, which includes an index of the term in question, its frequencies and the surrounding text, which can be sorted in one way or another [59]. It is worth noting that this kind of analysis is different from fulltext search. Concordance analysis "*is intended for understanding properties of language or for analyzing the structure and content of a document for its own sake, rather than search*" [59]. It helps users investigate terms' frequencies, study how they are used, gain insights on their evolving dynamics in meaning, or learn which words tend to go together in a text collection.

Unsurprisingly, concordance analysis is widely used in the literature or digital humanities domain, where an essence of the art is the focus on meaning. As we can think of "*meaning as residing in the distribution of contexts over which words and utterances are used*" [87], usage contexts play a key role in helping us understand how a term is used. Furthermore, it is contexts that help to deal with language phenomena such as polysemy (terms having multiple meanings) or synonymy (multiple terms having the same meaning). Not only is concordance analysis used in literature analysis, it is also employed in many other domains. For instance, in market analysis, it can be used to track how customers' responses to a product evolve over time; in investigative journalism, it can provide help in understanding the dynamics of the wordings used in a political context.

In our work on the IVEA application to support visual exploration of text collections described thus far, understanding usage contexts of a term requires reading the full text surrounding that seed term (by clicking on a cell in an Entities Distribution View described in Chapter 6 to jump to a text fragment containing that term). As this necessary reading is time consuming and requires a lot of effort from users in the case that there are many occurrences of the term, in this chapter, we focus on content abstraction for visual concordance analysis. This is a visual analytics challenge, which requires research in text analytics and information visualization. In Section 7.2, we discuss existing works in the visual concordance analysis area and their limitations. In Section 7.3.1, we report on the text mining technique used and in Section 7.3.2 we describe in detail the iterative process of designing the proposed visualization including an initial design and user feedback on it, which leads to current visualization. We also report on a user study, discuss the findings from it in Section 7.4 and outline future work in Section 7.5.

7.2 Related Work

In Section 3.5, we have discussed in detail existing works in the literature on visual concordance analysis. Apart from those, our work here is also related to other text summarization systems to a certain extent, as their main goal is to "*construct a characterization of*

document content through significant reduction of the original document source.... and incorporate the set of extracted, topically indicative fragments into a coherent representation of document content" [9]. While the literature on text summarization is vast, it is worth noting some recent works that are close to ours. In [34], keywords representative of emails are selected, based on a topic model, to show the gist of the content. However, there is no visual interface to convey the aggregation of keywords to assist users in digesting the summary. The TIARA system [83, 153] provides an interactive visual interface to provide a summary of a text collection. It uses the ThemeRiver metaphor [53] to show the ebbs and flows of topics within a collection of documents over time. The difference between TIARA and Context Stamp is that TIARA focuses more at collection level and not on the usage contexts surrounding a seed term within a document. TIARA shows the topic dynamics using different sets of keywords that best characterize a text collection at different time points. In general, text summarization systems related to our work with respect to the goal of abstracting away the details and presenting only the gist. However, we do not focus on the summary of a document as a whole. We focus on visualizing the usage contexts of a seed term or phrase and enabling comparisons between those contexts.

While being useful in different ways, the existing works on visual concordance analysis discussed above tend to have at least one of the following limitations:

- The contexts within which a term is used are usually shown as either in their original textual form or in a word cloud based simply on frequencies, which means that considerable efforts may still be needed from users to analyze them. While state-of-the-art text mining techniques are available, few approaches take advantage of them for semantic analysis.

- The distribution of a term within parts of a document is often not shown. Hence, a lot of useful information regarding a term's usages is unavailable to users.

- An entire document is usually treated as a unit of analysis, which could obscure meaningful interpretations of contextual information.

These limitations motivate the work reported in this chapter. When analyzing a set of

documents, e.g., the State of the Union addresses[1], one may ask the following questions:

- Was "*research*" mentioned early/late in or through out the speech in 2007? How frequently was it mentioned?
- What were the words used around "*research*" in 2007?
- How are they compared to those used in 2011?

Given the amount of text to analyze, these questions are best answered with a visual text analytics solution. However, the inherent categorical nature of text and its very high dimensionality make it very challenging to display the contexts graphically [59]. While the first question can be answered with a Seesoft-based visualization, such as the New York Times one mentioned earlier, the other two require further research into a visual representation that can show the gist of a term's contexts.

Our goal is to propose a novel approach to make it easy for users to quickly compare the sets of contexts within which a term is used in one document versus another. Here we consider:

- Instead of presenting a term's set of contexts in their original textual form, how can the details be abstracted away and only their gist retained to let users make contextual sense of the term?
- Which visualization elements can be used together to convey both the distribution of a term and its contexts at different levels of detail?

A solution to these problems will potentially provide support for users in sifting through information contained in term-bearing paragraphs and decide whether their contents are worth the effort of a deep read. Instead of focusing on the patterns or frequencies of text spans involving the seed term in question, we focus on analyzing its surrounding text. The proposed approach is discussed next.

[1] `http://www.presidency.ucsb.edu/sou.php` (Last accessed Jan 12, 2011)

7.3 Proposed Approach

As with other visual analytics solutions, our focus is not to propose a new visual metaphor, but to identify a good automated algorithm for the (text) analysis task, and then integrate the results with appropriate visualization and interaction techniques [70]. In the following sections, we describe how we employ a state-of-the-art text mining algorithm and visually transform its outputs to display complex dimensions of a term's usage contexts in an intuitive manner.

7.3.1 Text Analysis

It is worth pointing out that the approaches we employ thus far in our work have been based on a user-defined ontology encapsulating entities of interest. This ontology helps tailor the exploration toward users' interest, e.g., filtering for documents containing selected entities, or understanding their distributions within a document. However, this user-defined ontology does not tell us how the terms used in a particular collection are thematically related. These inherently data-driven thematic relationships are necessary for us to abstract away textual details of a set of contexts and yet retain facets of rich information about them. Therefore, in complement to the ontology-based approaches employed thus far, here we rely on a statistical topic model to obtain the relationship between the mental representation of language (meaning) and its manifestation in written form. Here we describe topic models and our approach for content abstraction.

Topic Models

Instead of treating documents as bags of words, a topic model treats documents as mixtures of latent topics, and each topic is a probability distribution over words [49]. A word can be assigned to various topics with different probabilities, depending on its levels of association with those strands of meaning. Unsupervised learning methods can then be used to learn the unobserved mixing coefficients for each document and the topic-word distributions.

Before describing the formal details of topic models, here we give an oversimplified example[2] of a topic model. Suppose we have a collection of short documents, 5 of which are about two topics, one is about air travel (topic 1), the other is about electronic devices (topic 2). The example documents are as follows:

- Document 1: "*The plane is about to take off.*"
- Document 2: "*This flight is delayed for two hours.*"
- Document 3: "*The new iPad has been released.*"
- Document 4: "*This laptop is very durable.*"
- Document 5: "*Only when I got to the airport did I realize that I did not have my iPad with me.*"

For these documents, a topic modeling technique should be able to infer that documents 1 and 2 are 100% about topic 1 (air travel), while documents 3 and 4 are 100% about topic 2 (electronic devices). Document 5, however, is 50% about topic 1 and 50% about topic 2. The technique would also infer that topic 1 consists of the following words (in a decreasing order of probability of belonging to this topic): "*plane*" (30%), "*flight*" (30%), "*pilot*" (25%), "*delay*" (15%). Meanwhile, topic 2 can consist of the following words: "*laptop*" (38%), "*computer*" (35%), "*iPad*" (12%), "*iPhone*" (15%). Note that the sum of the percentage of topics for each document and the sum of the probabilities of words for each topic are both 1 as they are probability distributions. The challenge here is the statistical process to infer these document-topic and topic-word distributions given a collection of documents.

To make this chapter self-explanatory, we hereby include a recap on key aspects of topic models from the seminal work reported in [8] and [49]; interested readers are referred to those publications for more details. Given a corpus consisting of **D** documents, a vocabulary of **W**, and a set of **T** topics, the probability of the *i*th word in a document is:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{7.1}$$

[2]The example is adapted from `http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/` - Last accessed date: 9 May 2012

whereby z_i is the latent topic variable from which the word was drawn, $P(w_i|z_i = j)$ is the probability of the word w_i under the jth topic and $P(z_i = j)$ is the probability of choosing a word from the jth topic in the current document. Here $P(w|z)$ indicates the importance of words in a topic and $P(z)$ shows the prevalence of topics within a document [49].

Let ϕ be the set of T multinomial distributions over the W words representing $P(w|z)$, and θ be the set of D multinomial distributions over T topics representing $P(z)$. To identify the set of topics, we need to obtain an estimate of ϕ that gives high probability (as in Eq.7.1) to the words within the corpus. The challenging nature of this task leads to the advent of Latent Dirichlet Allocation (LDA) [8], which is a complete generative topic model in that it incorporates prior probability distributions on θ and ϕ to allow for inference on unseen documents. As the Dirichlet distribution is conjugate to the multinomial, it is used as the priors on the multinomial distributions θ and ϕ.

The complete probability model of LDA is:

$$w_i|z_i, \phi^{(z_i)} \sim Discrete(\phi^{(z_i)})$$

$$\phi \sim Dirichlet(\beta)$$

$$z_i|\theta^{(d_i)} \sim Discrete(\theta^{(d_i)})$$

$$\theta \sim Dirichlet(\alpha)$$

Words and documents are generated from an LDA model as follows:

- For each document, pick a distribution over topics θ from a Dirichlet distribution.
- For each word within the current document, pick a topic j from the above distribution and then pick a word from that topic.

Given the observed data (words and documents), the inference task is to reverse the generative process to estimate θ and ϕ. An efficient inference method using Markov chain Monte Carlo was proposed in [49] and has been widely used to estimate document-topic and topic-word distributions. The Markov chain Monte Carlo procedure allows for obtaining samples from a complex distribution by letting a Markov chain converge to the target distribution and

then drawing samples from it. In particular, collapsed Gibbs sampling can be used, in which the next state of the chain is reached by sequentially sampling all variables from their distribution, conditioned on the current values of all other variables and the data. For LDA, only the latent topic variables $\mathbf{z}$ are sampled, based on the full conditional distribution $P(z_i|\mathbf{z}_{-i},\mathbf{w})$ [49]:

$$P(z_i|\mathbf{z}_{-i},\mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta}\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha} \tag{7.2}$$

in which $n_{-i,j}^{(w_i)}$ is the number of words assigned to topic j that are the same as w_i, $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, $n_{-i,j}^{(d_i)}$ is the number of words from document d_i assigned to topic j, and $n_{-i,\cdot}^{(d_i)}$ is the total number of words in document d_i, all not including the current word.

After a burn-in period, the chain approaches the target distribution, and the current values of z_i are used as samples. Given a sample $\mathbf{z}$, we can estimate θ and ϕ with:

$$\hat{\phi}_w^{(j)} = \frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+W\beta} \tag{7.3}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+T\alpha} \tag{7.4}$$

where $n_j^{(w)}$ is the number of times word w has been assigned to topic j, $n_j^{(\cdot)}$ is the total number of words assigned to topic j, $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j, and $n_{\cdot}^{(d)}$ is the total number of words in document d.

An efficient implementation for topic model inference using a data structure and scalable algorithm called SparseLDA is reported in [158], interested readers are referred there for technical details. The implementation is publicly available in the MALLET toolkit[3] [88], and is used in our work for fitting an LDA topic model for the sample dataset.

[3] `http://mallet.cs.umass.edu/` (Last accessed Jan 12, 2011)

Topic-based Content Abstraction

Here we use the set of State of the Union addresses given by US presidents from 1790 to 2011 to illustrate our topic-based content abstraction approach. The reasons for using this data set are:

- The data set is publicly available[4].
- The speeches are targeted at a general audience and hence do not use any technical language or jargons.
- Each speech covers a wide range of topics.
- The inherent temporal aspects of the speeches can enable users to make interesting comparisons about matters discussed around a seed term at different points in time.

The only pre-processing we did was to remove stop-words, as well as annotations such as "*[laughter]*", "*[applause]*" from the original transcripts. A model with 120 latent topics (the number of topics needs to be defined in advance to adjust the latent topics' granularity) is built for this dataset[5].

With this model, we have the inferred distributions of topics within documents, and the distributions of words over the 120 topics. The key outcomes are the *decompositions of the inferred topics*, which are coherent clusters of thematically related words, and the *assignments of topics to words* in documents. For instance, in this model, two example inferred topics consist of the following thematically related words, in decreasing order of probability of belonging to the corresponding topic:

- Topic 17: "*jobs*", "*tax*", "*american*", "*plan*", "*spending*", "*reform*", "*budget*", "*bill*", "*long*", "*pay*", "*make*", "*put*", "*business*", "*rates*",...
- Topic 115: "*bank*", "*banks*", "*money*", "*notes*", "*money*", "*institutions*", "*banking*", "*prices*", "*deposit*", "*moneys*", "*required*", "*public*", "*deposits*", "*revenue*","*credit*",...

[4] `http://www.presidency.ucsb.edu/sou.php` (Last accessed Jan 12, 2011)

[5] Note that the resulting model is slightly different from the one initially reported in [137] as there is a new addition of the 2011 speech, the latest at the time of writing.

In addition, the topic inference process also results in each word in each document being assigned a topic. Figure 7.1 shows an example paragraph in the 2007 speech, which is annotated with topics assigned to words. Within this paragraph, words such as "*vital*", "*changing*", "*generates*", "*electric*", "*diesel*", "*ethanol*" were assigned to the same topic with different probabilities.

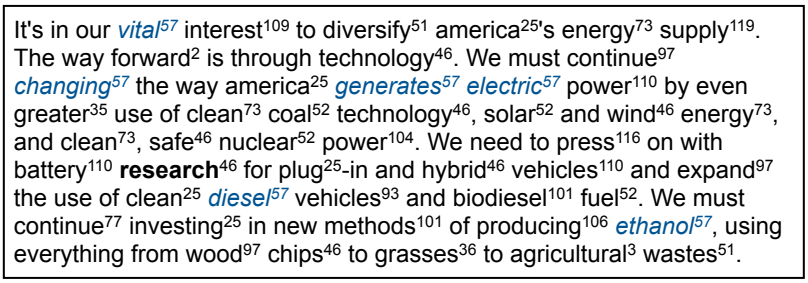


Figure 7.1: A paragraph with topics assigned to words. Words belonging to the same topic 57, such as "*vital*", "*changing*", "*generates*", "*electric*", "*diesel*", "*ethanol*", are in italic.

While the inferred model is imperfect as finding the optimal model parameters for a dataset is non-trivial [149], these topic-word distributions can be employed to abstract away textual details of a term's set of contexts in a document with the following steps:

1. Identify paragraphs containing the seed term[6], together with their lengths and the locations of the term's occurrences.

2. For each paragraph, iterate over all words, get their assigned topics and put words belonging to the same topics into "topical" groups.

3. The total frequency of all words within the group is used to visualize its size. In each group, select its representative word. As stop-words have been removed, the most frequent word acts as the representative of a group. Tie-breaking[7] is based on a weighted score that is proposed in [7]. This score takes into account both the per-topic probability (the importance of a word within a topic) and whether a word has high probabilities under many topics (the score is down-weighted if this is the case). The

[6]Without loss of generality, a term can be either a single word or multiple words. In the latter case, all words need to appear together to be considered as a match.

[7]Tie-breaking is only added in the latest version, not in ContextBurst and an early version of Context Stamp.

weighted score for word w in topic j can be calculated as follows [7]:

$$weightedScore = \hat{\phi}_w^{(j)} log(\frac{\hat{\phi}_w^{(j)}}{(\prod_{t=1}^{T} \hat{\phi}_w^{(t)})^{\frac{1}{T}}}) \tag{7.5}$$

4. Similar to steps 2 & 3, but aggregate the data over all term-bearing paragraphs to obtain the data at document level[8].

Paragraphs are used as units of analysis instead of sentences as in most related work because: (1) written text tends to be structured such that each paragraph encompasses a different focus; (2) sentences in a paragraph that appear before or after a term-bearing sentence may refer to that term using co-references. Hence, retaining only term-bearing sentences would potentially discard a lot of interesting information.

7.3.2 Visualization Design

In order to assist users in consuming the outputs from the text analysis stage, we need to design a visualization that is able to convey the concordance of a term well. As visual concordance analysis tasks can be done by a broad base of users and not just a specific group of experts, it is important to strike a balance between aesthetic and functional properties of the target visualization. Similar to other user interface designs, the visualization of contextual information goes through an iterative process. In this section, we first describe the initial design of the visualization and the feedback solicited from users in informal interviews, then go into detail of the current alternative.

Early Version

The topical groups of context words surrounding the seed term are visualized by ContextBurst, the initial visualization that is based on SunBurst [127] and DocuBurst [22]. Both are space-filling visualizations that use a radial layout. Figure 7.2 shows a ContextBurst visualization displaying the usage contexts of the seed term "*insurance*" in the State of the

[8]Note that this approach is to produce visualization of a term's concordance, not of the distribution of topics within a document as a whole. It only takes into account those paragraphs that contain the seed term.

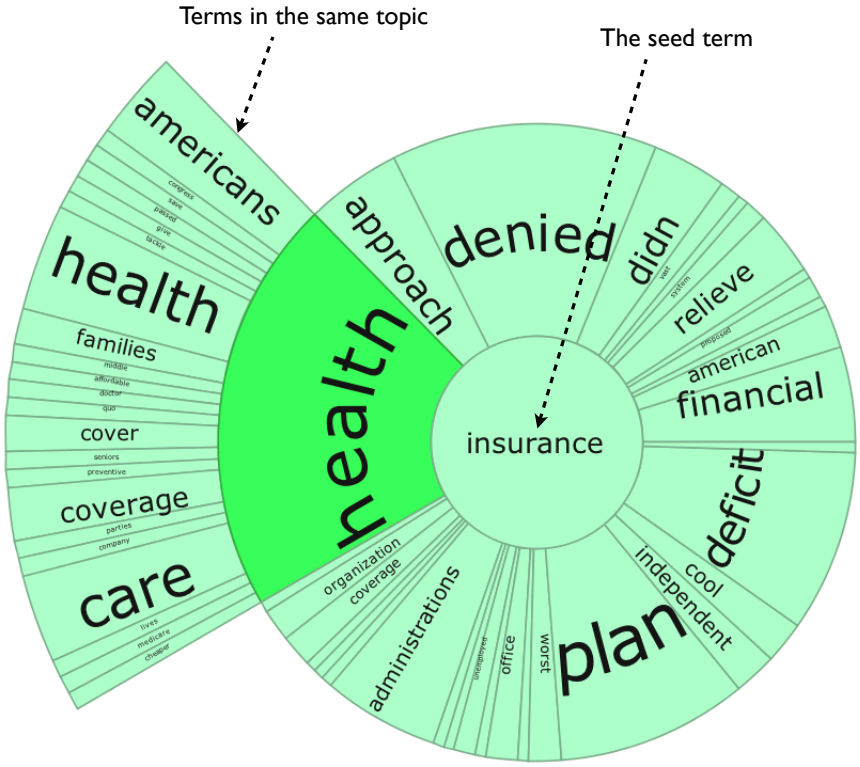


Figure 7.2: ContextBurst - the initial visualization used to display usage contexts of the seed term "*insurance*" within the 2010 State of the Union speech.

Union address in 2010. In the ContextBurst visualization, the seed term is put within the inner circle at the center, topical groups identified in step 4 of the approach outlined in Section 7.3.1 are placed in the surrounding wedges at the next level. The angular widths of the wedges are proportional to the number of words these groups contain. At first, only the representative word in each topical group is shown in the corresponding wedge. When users click on a wedge, the list of words within that topical group are then shown at the second level. Text labels are displayed with the largest possible sizes to enhance legibility. Figure 7.2 shows the decompositions of different topics discussed in paragraphs containing "*insurance*" in the 2010 speech. When users click on the topical group represented by "*health*", they can see all other terms belonging to the same topic, such as "*americans*", "*care*","*coverage*","*cover*","*family*", etc. We were interested in using this existing visual metaphor to enable users to interact with the results of the text analysis because of its visual structure, which allows for the seed term to be placed in the center, and then surrounded by

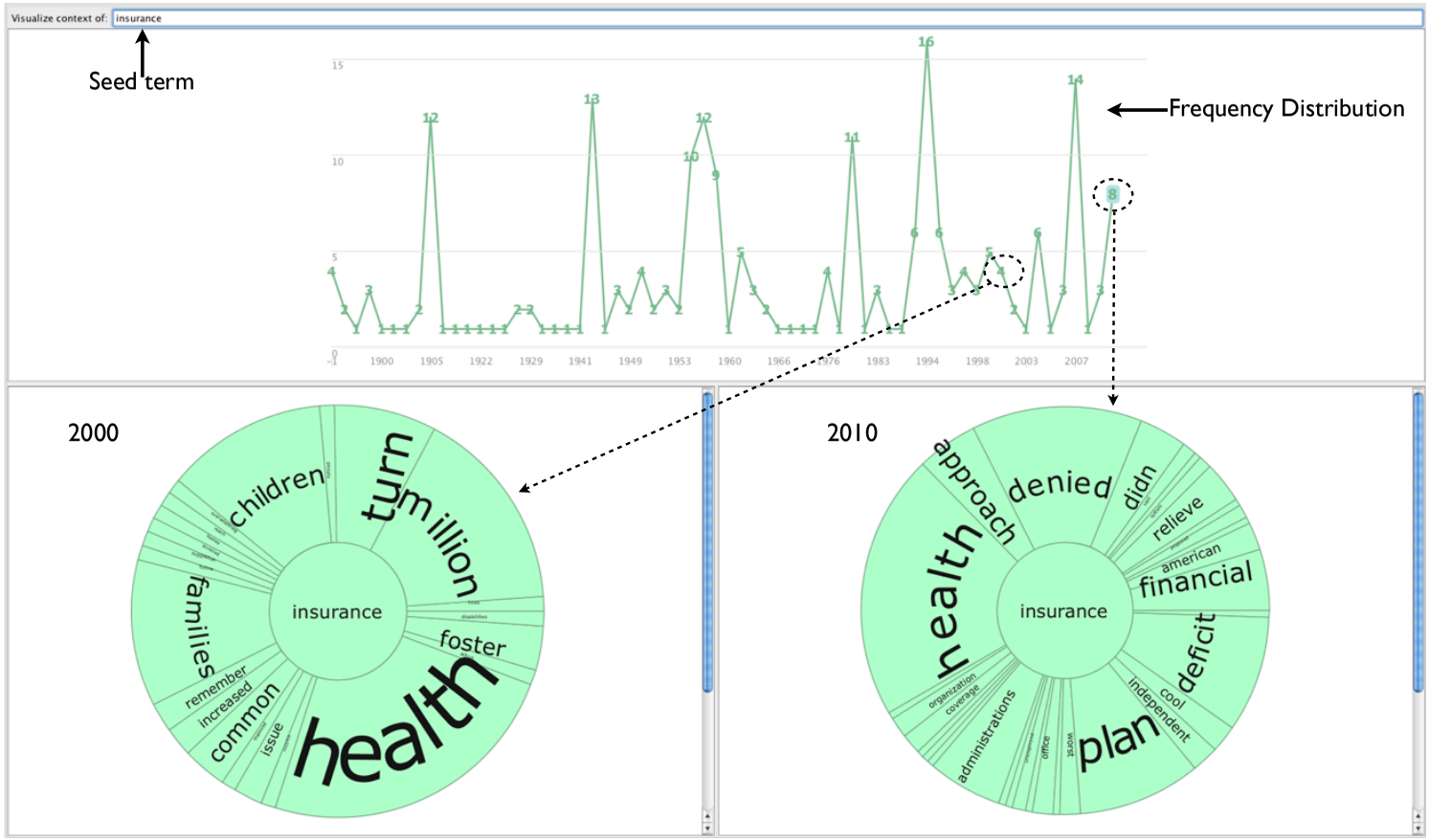


Figure 7.3: ContextBurst - Comparison of usage contexts of the seed term "*insurance*" within the 2000 vs. the 2010 speech.

words representing different topical groups. This visual form per se appears to convey the nature of usage contexts well.

Furthermore, in order to enable comparisons of usage contexts of a term at different time points, we integrate a timeline to show the frequencies of occurrence of the seed term over time as shown in Figure 7.3. When users click on a dot on the timeline, the application will either (1) show a list of documents having the selected year attribute in the case that there are multiple documents in the selected year, or (2) show the contextual information of the document having the selected year attribute if the dataset contains only one document for that year. Figure 7.3 illustrates the use of ContextBurst to compare usage contexts of the seed term "*insurance*" in the year 2000 versus 2010. The implementation of ContextBurst is based on the prefuse toolkit [61] and the source code made publicly available by DocuBurst's author[9].

With ContextBurst as the initial attempt at visually depicting the text analysis outcomes, we were interested in its advantages and disadvantages or potential usability issues when used to show the usage contexts of a term. We conducted informal interviews with six people to solicit their feedback. We explained the purpose of the prototype and then let the participants freely interacted with ContextBurst on the State of the Union dataset. The subjects then reflected on what they liked and what they did not like about the visualization.

The feedback was very positive with respect to the usefulness of this kind of visual concordance analysis application. With the timeline alone, the subjects appreciated that the easily visible peaks and valleys of a seed term's frequencies already enabled them to infer some information or hypothesize about certain events. In addition, ContextBurst provided an overview of a term's concordance, and the ability to see two ContextBursts side by side for comparison was well-liked by the participants.

However, ContextBurst was also thought to have certain usability drawbacks:

- As ContextBurst only showed a term's concordance at document level, there were too many words in the same topic that are visually presented at the second layer.

- Five out of six subjects thought that a view showing the distribution of a term within a

[9] `http://faculty.uoit.ca/collins/research/docuburst/index.html` (Last accessed Jan 12, 2011).

document would be useful.

- Some labels were "*very hard to read*" considering the different angles and forms (i.e. either in a straight line or along a curve, which was intended to maximize their font sizes).

- It might be better to use different colors for different topical groups. In ContextBurst, the lack of visual cues to identify topics made it difficult to relate how different words in the same topic were used in comparison between documents.

- When a term is mentioned in many documents (e.g., "*education*"), i.e. many years in the case of the State of the Union data set, the full timeline was too tightly packed.

Current Design

Given the shortcomings of ContextBurst for concordance analysis tasks, we considered an alternative design. To show both the distribution of a term and the decompositions of different aspects of its surrounding contexts, we propose Context Stamp, a visualization that is an innovative combination of the TreeMaps metaphor [119] and the Seesoft visualization [35]. With "*invest*" being the seed term, Figure 7.4 shows two Context Stamps at paragraph level on the speech given in 2006, which consists of two parts.

The first part is the *term distribution view* consisting of a set of vertically stacked bars (on the right side of Figure 7.4) representing paragraphs within a document[10]. The bars' heights are proportional to the lengths of their corresponding paragraphs. Occurrences of the seed term are represented as round dots, with their positions within the bars corresponding relatively to where in a paragraph they appear. For instance, the term distribution view in Figure 7.4 shows that the term "*invest*" appears once in two paragraphs in the speech given in 2006. These dots serve as an index into the documents, as contents of term-bearing paragraphs are shown, with the term's occurrences highlighted, when users click on the dots. Therefore, users can have instant access to the original content of paragraphs in full.

[10]Note that while the TileBars-based Entities Distribution View discussed in Chapter 6 can also be a good candidate here, we did not choose to use it because we want to show the term distribution at a more fine-grained level. In the TileBars-based Entities Distribution View, we cannot show occurrences of terms individually but only their frequency within each fragment of a document.

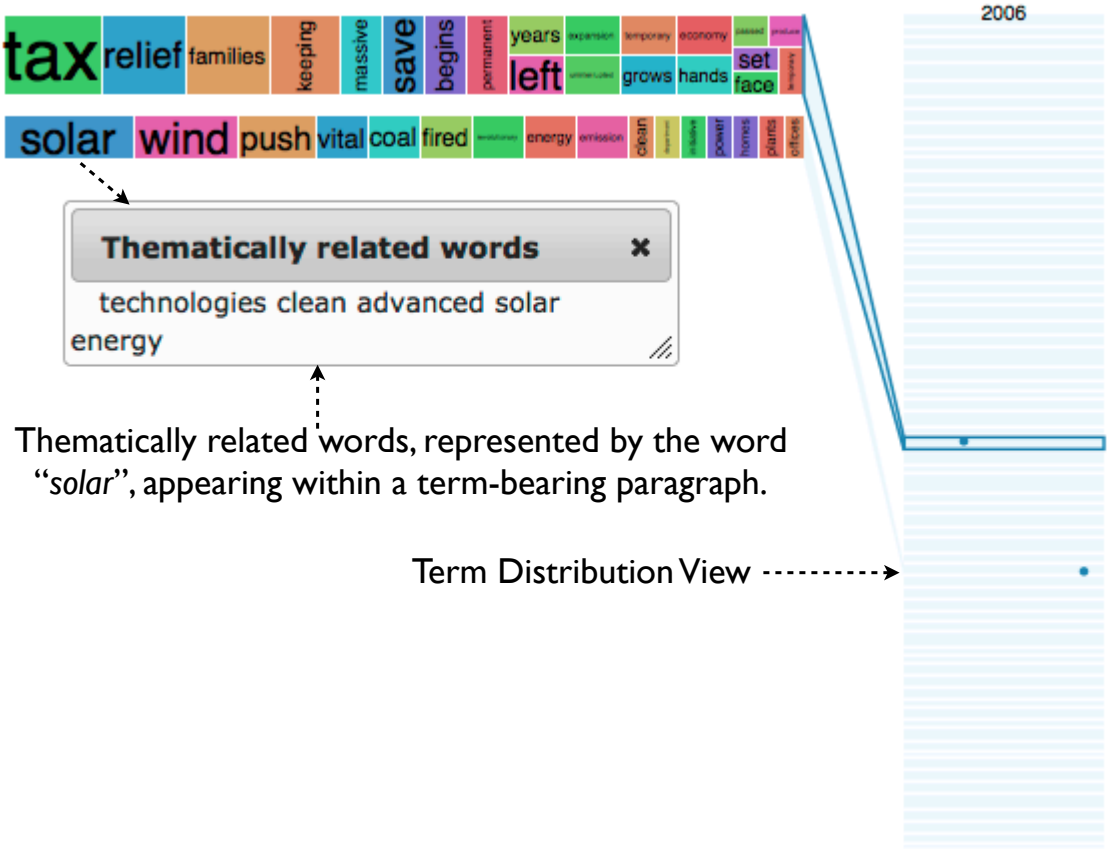


Figure 7.4: Small multiple Context Stamps at paragraph level with the seed term "*invest*" in the speech given in 2006. Note that the seed term is not shown in this visualization, only in the search box of the application (see the top part of Figure 7.5).

With this visualization alone, in many cases the seed terms' frequencies and distributions might help users hypothesize about their importance within a document. For instance, terms mentioned more often or earlier within a document might suggest a higher importance or more urgency.

The second part is a set of small squarified TreeMaps, called *Context Stamps*, which represent the gist of the contexts in each term-bearing paragraph. The gist is constructed based on the outputs from Step 3 in the text analysis process described in Section 7.3.1, i.e. instead of presenting the full text of a term-bearing paragraph, we only show the set of topics assigned to words within that paragraph. Each topic is shown via its representative word, and its size is the number of words assigned to that topic. At paragraph level, there is a Context

Stamp for each term-bearing paragraph. The design decision on using a rectangular space-filling representation is to accommodate for displaying many such visual elements in parallel with the term distribution view. The TreeMaps metaphor is a good fit for this purpose, as it uses all the allotted space efficiently, and this is not the case with ContextBurst (it is not possible to assign one ContextBurst to each term-bearing paragraph as that would require too much space). In this visual form, we also opt to not show all words within a topic in child squares inside the topic's parent TreeMaps cell. This design choice is made for two reasons: (1) to meet the goal of abstracting away the details and only retaining important elements of the text and (2) even if they are displayed in subdued colors or with high transparency, they would still result in clutters of text within TreeMaps cells that are of limited sizes (note: we are using multiple small TreeMaps stacked up on each other instead of a single big one as how TreeMaps are typically used). Therefore, each topic is labeled by its representative word (within the paragraph scope), which occupies a cell of the paragraph's TreeMaps. The cell's area size is determined by the total frequency of all words belonging to that topic in a paragraph. Each topic is assigned a color to enable visual cross-comparisons. To enhance the legibility of words without overflowing the containing cells, font sizes of representative words within TreeMaps' cells are dynamically calculated based on the cells' areas such that the font sizes are the largest possible. Clicking on a TreeMaps cell will pop up a list of thematically related words in full. Figure 7.4 shows this interaction on a topic represented by the word "*solar*". In addition, there are also visual cues that link between the TreeMaps and their corresponding bars (term-bearing paragraphs) upon mouseover.

By hiding away words that are thematically related, in effect Context Stamps retain the most important topical dimensions using their representative words, and at the same time abstract away much information that would have required users' perusal of free text. As a result, paragraph level Context Stamps can offer an immediate overview of the main aspects of term-bearing paragraphs within a document, i.e in the speech given in 2006, the first paragraph containing the term "*invest*" is mainly about tax, reliefs, families matters, while the second paragraph mentioning it is about different kinds of energy such as solar, wind and coal. Via its space divisions, Context Stamps also helps users in getting a quick idea of the most heavily discussed aspects as well as the least mentioned ones within these paragraphs.

Apart from serving as small multiple TreeMaps at paragraph level, Context Stamps can also be used at document level. This is especially beneficial for high level comparisons of a seed term's usage contexts within one document versus another or at one time point versus another. In Figure 7.5, a focus+context timeline[11] shows the frequencies of occurrence of the seed term "*research*" in the speeches given over the years. When users click on the points on the timeline, the resulting Context Stamps for the selected years are shown side-by-side. Here, a single Context Stamp is used to visually depict the aggregated data obtained in Step 4 of the content abstraction approach detailed in Section 7.3.1. It is worth emphasizing that this is about visual concordance analysis, not visualization of topic models per se. Therefore, the topic decompositions within a document level Context Stamp only correspond to the topics identified within term-bearing paragraphs and their aggregated weights, and they do not reflect the decomposition of all topics for the whole text in a document. As shown in Figure 7.5, document level Context Stamps can facilitate comparisons of not only the term's frequencies and distributions within two documents, but also the main topics, together with their proportions, that were discussed around "*research*" in the speech given in 2007 versus the one in 2011. The difference can easily be inferred or hypothesized from Figure 7.5 without much effort from users.

In sum, our proposed approach toward content abstraction in visual concordance analysis provides visual perspectives on a text that are potentially helpful in many ways, as they can act as visual summaries of term-bearing paragraphs and subsequently provide navigational pointers for close reading.

In terms of implementation, the visualization is written in JavaScript based on the Protovis toolkit [10]. The web application prototype is developed using the Google Web Toolkit, running on an Apache Tomcat server.

[11] The focus+context timeline implementation is adapted from the Protovis example gallery at `http://vis.stanford.edu/protovis/ex/` (Last accessed Jan 12, 2011).

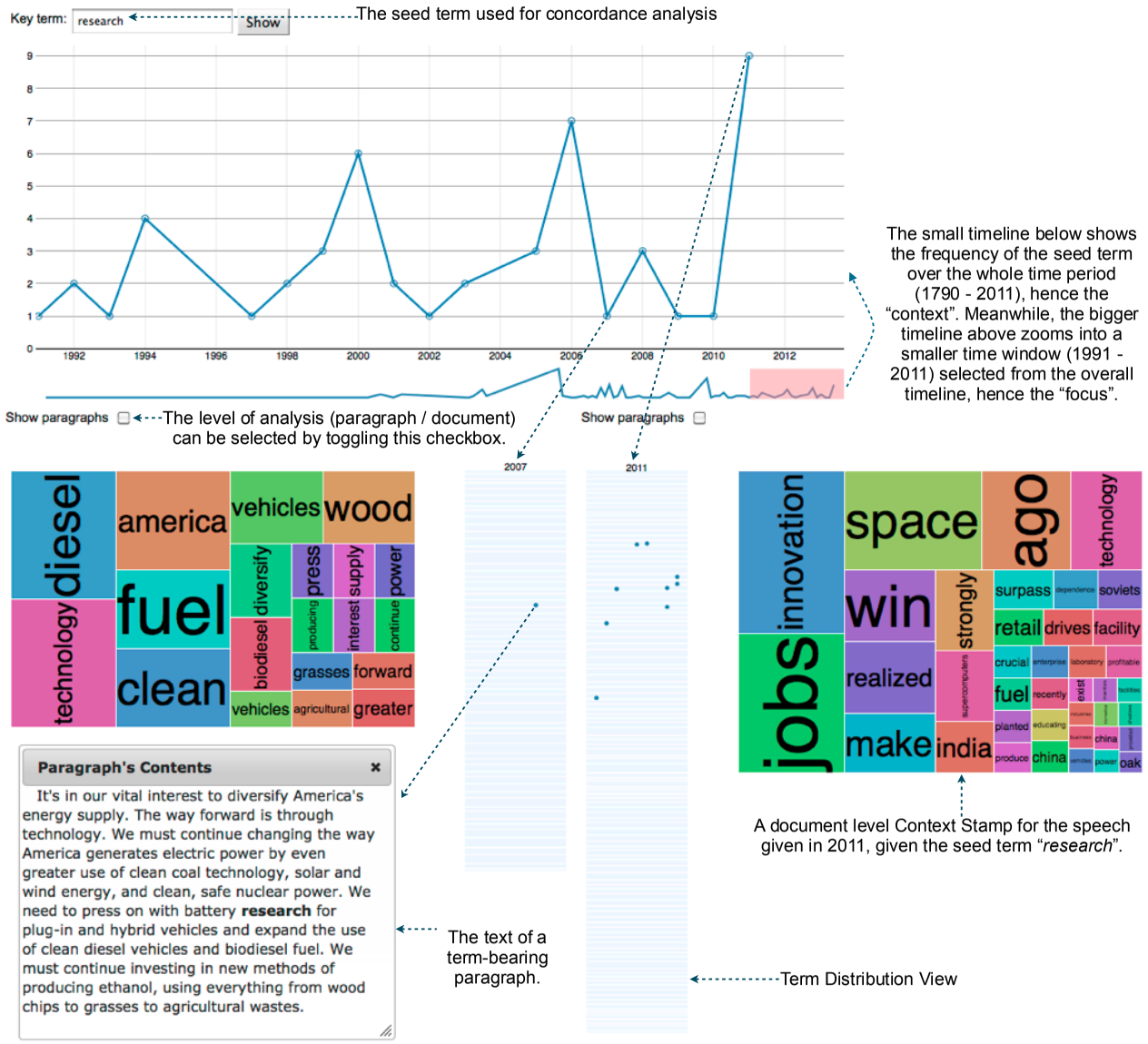


Figure 7.5: Context Stamps at document level can be used for comparison of contexts across documents in 2007 vs. 2011.

7.4 Evaluation

7.4.1 Text Analysis

The topic model is evidently at the center of our proposed approach. Therefore, while it is not our focus here to carry out research on approaches toward improving the quality of topic models, it is important for Context Stamp visualizations that the trained model be a suitable one.

There are many factors involved in the inference process of an LDA model: the hyperparameters α, β, and the number of topics T. Despite its introduction from many years ago, LDA is still considered to be "*still in a relatively early stage of development*" [154], and determining the optimal values for these parameters in order to train a topic model given a corpus remains a research challenge. While the choice of the hyperparameters α and β can have important implications on the resulting model [49], most researchers typically use symmetric Dirichlet priors with heuristically set values for α and β. When it comes to selecting the number of topics T, given the values of α and β, the traditional approach is to compute the posterior probability of held-out documents on a set of models with different values of T and choose the best one [49].

A more recent study reported in [149] shows that (1) using an asymmetric Dirichlet over the document-topic distributions and a symmetric Dirichlet over the topic-word distributions results in significantly better model performance and (2) using optimized Dirichlet hyperparameters results in improved consistency in topic usage as T is increased, which means that if LDA has a sufficient number of topics to model the observed data well, a larger number of topics would not significantly affect inferred topic assignments [149]. Therefore, in our work we apply these settings accordingly to achieve a suitable topic model for the State of the Union dataset. The impact of the resulting model is discussed later in Section 7.4.3.

7.4.2 Visualization

We carried out an evaluation to gauge Context Stamps' effectiveness in supporting users in comparing a term's contexts. The method is as follows.

Participants

Fourteen people (ten females, four males) participated in the study, one of whom was in a pilot session. Six of them were in the age range of 18-25, seven in 26-40, and one in 41-60. In terms of occupation, the subjects included a lecturer, a research associate, software engineers / business analysts, undergraduate and postgraduate students. All were familiar with web search and had a need to seek for information from documents (with frequencies ranging from a few times a week to hourly). None had any prior experience with the Context Stamp prototype.

Materials

We used the set of publicly available State of the Union speeches from 1790 to 2011 as the test collection in the study. While this is not a large collection in terms of the number of documents, each speech is of great length. More importantly, the wordings used in these speeches are of significant interest (demonstrated by a dedicated visualization by the New York Times mentioned earlier) as they were carefully chosen to convey the intended meanings, and their time stamps carry an important role in the concordance analysis task.

Design

In this study, we implemented three different interfaces to display the contexts in which a seed term is mentioned. The visual elements that are commonly available in all three versions are the timeline and the term distribution visualization. All three versions are available in the same webpage, selection of interface can be done simply via a radio button click.

- Fulltext-based interface: This baseline interface resembles the visualization provided by the New York Times mentioned earlier. The main difference between the Fulltext interface and that visualization is that the New York Times visualization only provides users with a single paragraph upon a mouse-click on a dot in the term distribution view. With the Fulltext interface, when users click on a dot in the timeline, all paragraphs containing the seed term are extracted, and shown together linearly in their full text

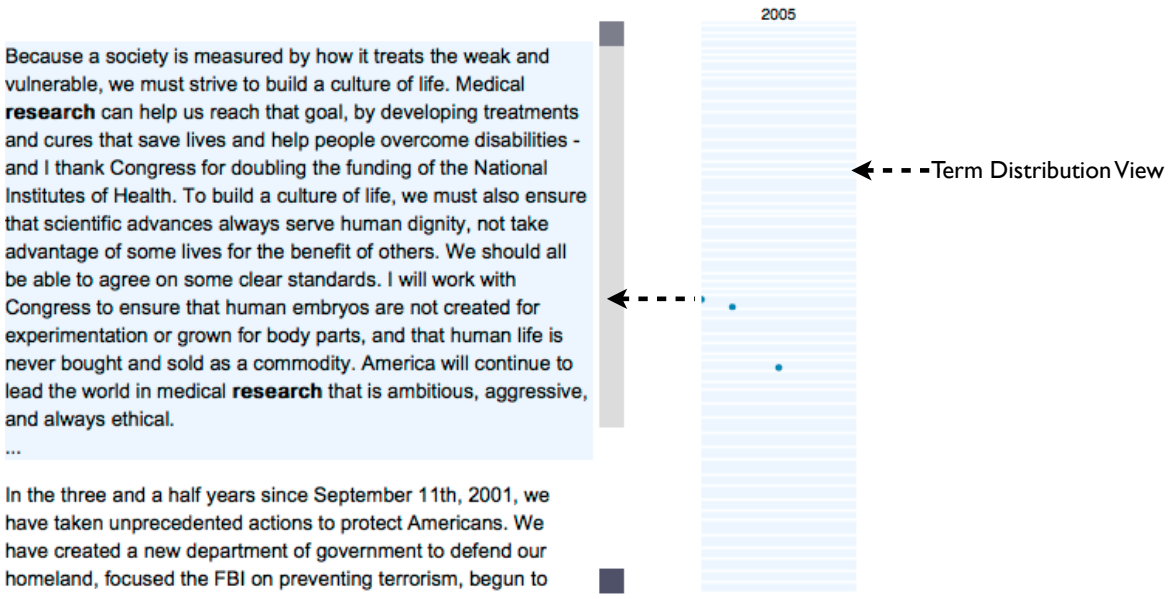


Figure 7.6: The fulltext display of the contexts of the seed term "*research*" within the 2005 State of the Union speech.

form, with the seed term displayed in bold. When users click on a dot in the term distribution view, the corresponding paragraph is then scrolled into view (in case there are many such term-bearing paragraphs), with its background color changed to light blue to distinguish it from others. Side-by-side comparisons are possible as in the Context Stamp visualization. Figure 7.6 shows an example of the Fulltext interface with the seed term being "*research*" and the speech given in 2005. Note that there is no separation of paragraph and document levels in this interface as all term-bearing paragraphs are aggregated in one place and a paragraph in focus is interactively highlighted.

- Word cloud-based interface: This interface employs a feature that is common in most related works in the visual concordance analysis literature, that is showing a frequency-based word cloud. Here stop-words are also removed from the original text. At paragraph level, each term-bearing paragraph is visually represented by a word cloud with the most frequent words placed in the center, as seen in Figure 7.7 for the seed term "*research*" within the 2003 State of the Union speech. At document level, we take into account all words in term-bearing paragraphs and show an aggregated frequency-based word cloud, as shown in Figure 7.8 for the same seed term and speech. We used

a freely available jQuery plugin[12] to generate these word clouds.

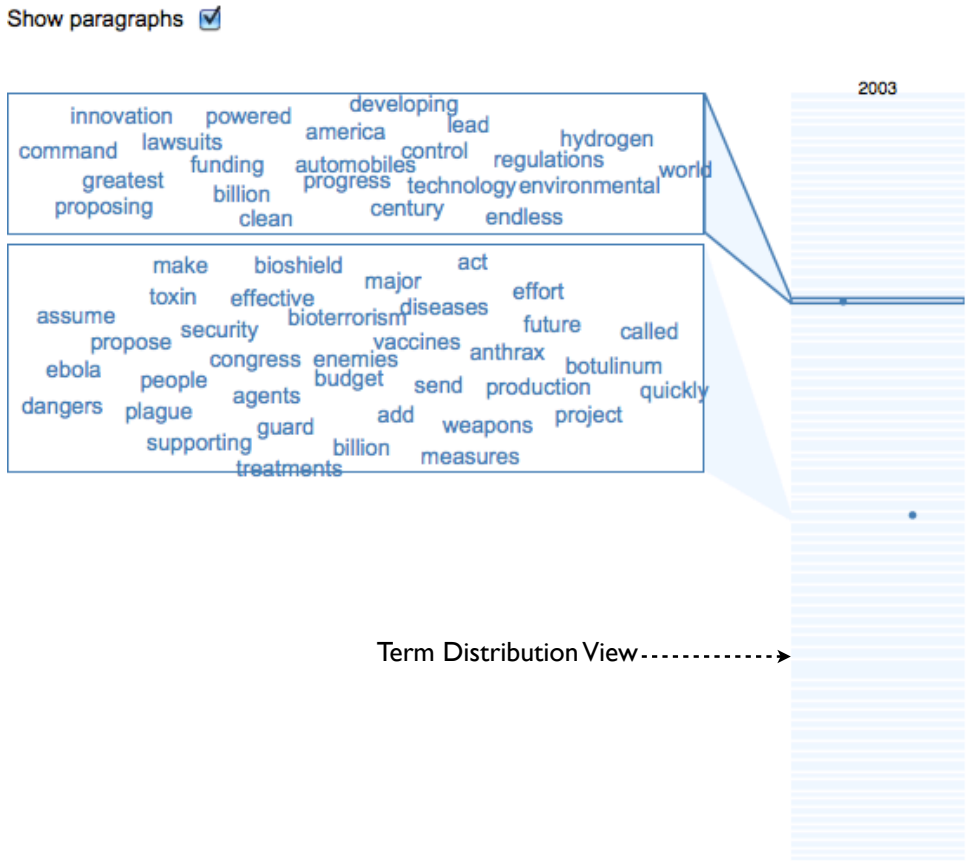


Figure 7.7: The word cloud display of the contexts of the seed term "*research*" within the 2003 State of the Union speech at paragraph level.

- Context Stamp: as described earlier in Section 7.3.2.

The above three interfaces were considered such that they enable the study participants to gauge how Context Stamp and different visual concordance analysis applications typically used in both the literature and the mainstream media. The participants were not told which interface was our proposed solution.

Given these interfaces, it may be tempting to run a controlled experimental study. However, considering the analytical nature of the concordance analysis tasks, controlled experiments are distant from their exploratory goals. Benchmark tasks used in controlled experiments are considered as not representing exploratory analysis tasks well, due to the following four characteristics, identified in [93]:

[12] `http://plugins.jquery.com/project/TagCloud` (Last accessed Aug 18, 2011)

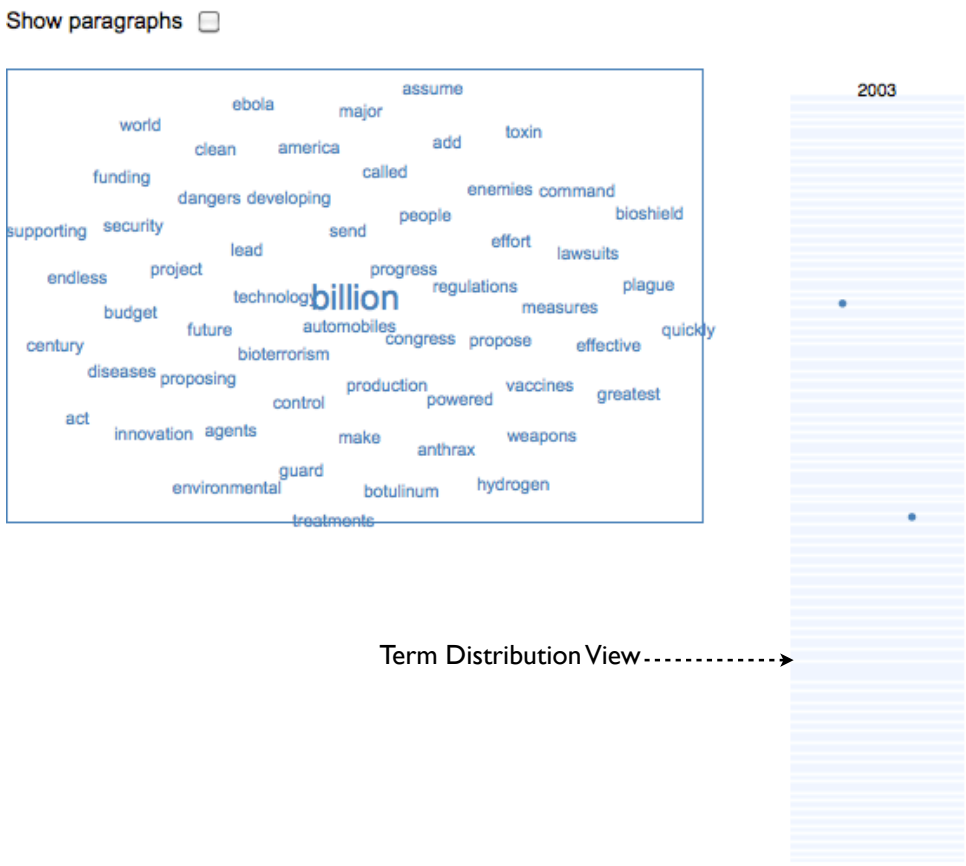


Figure 7.8: The word cloud display of the contexts of the seed term "*research*" within the 2003 State of the Union speech at document level.

- They must be predefined by test administrators and hence users must follow specific instructions.
- They need definitive completion times.
- They must have definitive answers that measure accuracy.
- They require answers that users can easily specify.

Therefore, controlled experiments would leave little room for users to freely explore the data collection. This calls for a suitable study involving the following (proposed in [93]):

- An open-ended protocol in which users explore the data in their own way.
- A qualitative insight analysis in which users report their findings and those are then coded / analyzed by the researchers.

- An emphasis on domain relevance whereby users may go beyond dry data analysis and make domain-specific inferences or hypotheses.

In [121], the "*Multi-dimensional In-depth Long-term Case studies*" (MILC) paradigm is proposed. This paradigm combines various methods including ethnographical participant observation methods, interviews, surveys, and automated logging of user activities. While comprehensive, this paradigm might be too costly for researchers [45]. In addition, discoveries can have a significant impact, but "*they occur very rarely, making it difficult - if not impossible - for someone to be observing when a discovery occurs*" [121].

As such, we conducted a user study using the diary technique, together with analysis of questionnaires, usage logs, and post-hoc interviews where possible. We did not log the tool usages by the participants individually on the client side. While it might be useful to have specific data such as which interface was used, the sequence and period of usage, it is hard to get the exact details. For instance, the time spent on a particular visualization might include users doing something else at their workplace that we would not be able to know of, and many people might have used the same computer, or the same user might have used different computers, etc. Therefore, at this scale, we only collectively logged the interface, the date, the seed terms issued, the years of documents selected for concordance analysis based on requests to the web server.

Procedure

The participants were invited to use the three interfaces over a period of at least one week. They accessed the web application from their own computers, at times and places of their convenience. There were no controls or influences by the researchers on the physical settings. In addition, there were no structured tasks, they could use any seed terms with any of the three interfaces they wished to. The only information they received from us is a link to the evaluation website which included an FAQ page that explained briefly what concordance analysis is and how they could use the three interfaces. Once the subjects were finished with exploring the text collection using the web application, they filled out a questionnaire, which asked for qualitative feedback with the following questions:

- Which combinations of seed terms, years and interfaces did you use that result in some findings? Please briefly explain your findings and classify each of them into one of the following categories: expected / interesting / very surprising.
- How do you rank the three interfaces, i.e. which interface helps you learn the most about usage contexts of terms in this collection?
- How do you rank the perceived learning curve over time for each interface? (1: easiest to learn, 3: take a while to get used to).
- On a scale from 1 to 9, how confident are you with the results shown in each interface?
- Would you use any of the three interfaces? If yes, please indicate which one and give an example of a domain of interest.
- What kinds of improvement would you like to suggest?

Apart from those questions, we also solicited their subjective ratings on the following adjectives for each of the interfaces: "*Easy to use*", "*Stimulating*", "*Satisfying*", "*Overwhelming*", "*Flexible*", "*Self-descriptive*", "*Organized*","*Tedious*". The ratings were on a Likert scale from 1 (Strongly disagree) to 9 (Strongly agree), with 5 being neutral. Similar to the study conducted in [159], we used a wide range to have a more sensitive testing instrument.

Results

Findings from the text collection The server log indicated that the Fulltext interface was used 452 times, the Word cloud interface was used 207 times and the Context Stamp interface was used 569 times. In their feedback, the participants reported some examples of the findings during their exploration of the text collection. In addition, they also indicated their subjective opinions on whether each of the findings was expected, interesting or surprising. The number of findings reported by the participants, grouped by type, is summarized in Table 7.1. Note that there are findings that were reportedly reached by using only the timeline, as well as findings that were encountered while the subjects were using all three interfaces.

Some examples of the findings are:

Type/UI	Fulltext	Word Cloud	Context Stamp	All three	Timeline only
Expected	5	5	7	3	0
Interesting	1	5	8	8	1
Surprising	0	2	10	0	3
Total	6	12	25	11	4

Table 7.1: Findings from the Text Collection by Type

- On "*Firefighter*": "*The first occurrence of this term in any of the speeches is in 2002. It is not mentioned before this year... I used the context stamps. From this I could tell that the 2011 content was related to a different matter than the 2002 content. A bunch of technology terms were in the 2011 paragraph while the 2002 paragraph was represented by terms relating to commemoration. I found both these things interesting because in the first case, I didn't think the term would not have been mentioned before 2002 and in the second case, I didn't expect this term to be mentioned in the context of internet connections. This is surprising. I also had a look at the word-cloud representation of the same two years as above. The terms on display were relevant to the issues at hand but the context-stamp seemed to highlight more useful words in a clearer, more prominent way than the word cloud like, for example in 2011, the word "wireless" is obvious in the context stamp representation but not in word cloud.*"

- On "*Research*": "*I was curious to see what were the main research focal points during the cold war(1981) and 2011. This was interesting for me since in the first paragraph of both years, the presidents talked about solar energy. Other commonalities between these two years were the research in space programs and the importance of education.*"

- On "*Education*": "*Seeing results between 2000 and 2011, presidents opted for tackling education in different ways. We see how education was used as a solution to criminality and drug reform (2000 - 2006). Also talks for educative reforms were tackled to decrease taxes and increase grants. Whilst in the last 4 years (2008 - 2011), education was a central part w.r.t. research and economy.*"

Rank	Fulltext	Word Cloud	Context Stamp
1	2	4	**7**
2	2	**5**	6
3	**9**	4	0

Table 7.2: Helpfulness Rankings

Helpfulness Table 7.2 shows the number of times the subjects ranked the three interfaces at each rank position in terms of their helpfulness. The Context Stamp interface was ranked as helping users learn the most about usage contexts of terms in the text collection by more than half of the participants. The Word cloud was ranked most often in the second position and the Fulltext interface in the third.

Typical comments on their helpfulness are:

- Fulltext: "*Even though the full text is displayed to the user, it can be quite difficult for him/her to detect the context of the terms within the collection if the paragraph is long and contains a lot of text. You also need to read all the paragraphs that contains a term that you are searching for, in-order to detect the context that the term is being used in.*"

- Word cloud: "*The Word Cloud doesn't seem to be quite as quick as the Context Stamp because there are many more terms contained in it. Also, it seems that sometimes the most prominently displayed words in the Word Cloud (the ones in bold) are the ones most mentioned in the paragraph but are not necessarily words which are key to the message of the paragraph.*"

- Context Stamp: "*Context Stamp allows me to skim more text more quickly, and focus on particular important terms.*", "*This type of interface helped me most in identifying the main themes which were discussed in the speeches.*"

Perceived learning curves Table 7.3 shows the number of times the subjects ranked the three interfaces at each position in terms of their perceived learning curves over time. There were three cases in which the Fulltext and the Word Cloud interfaces were ranked equally

Rank	Fulltext	Word Cloud	Context Stamp
1	**12**	2	0
2	2	**8**	4
3	1	0	**10**

Table 7.3: Perceived Learning Curves Rankings

first, and one case where the Word Cloud and Context Stamp interfaces were ranked equally second.

Some example comments on their perceived learning curves are:

- Fulltext: "*The fulltext interface is easiest to learn as it simply displays all of the content. Therefore, the context of the search term can be understood by reading through the text. It is like reading a newspaper.*"
- Word Cloud: "*Used to such tag clouds in blogs, thus it wasn't difficult to get used to*."
- Context Stamp: "*Context Stamp is straightforward but in some ways complex – because in order to make the sizes meaningful, sometimes the words are too small to read. By paragraph vs. overall is another difference I would have to see and understand. (Keep that, it's useful – just know that it does add complexity!)*" (Note: by "*paragraph vs. overall*" the subject meant paragraph level and document level.)

Confidence with results We were interested in testing the null hypothesis that the distribution of the confidence of the subjects on the presented results is the same across the three interfaces. A Friedman test was conducted and there was a statistically significant difference in the confidence with results, $\chi^2(2) = 11.617, p < .05$. Pairwise comparisons in post-hoc analysis, using Wilcoxon Signed-Rank tests at the Bonferroni-adjusted significance level of .05/3 = .017, showed that there was a statistically significant difference between the confidence on the Context Stamp Interface (median rank = 6) and the Fulltext Interface (median rank = 9) , with $Z = -2.605, p < .017$. There were no other statistically significant differences. A typical comment in this respect is: "*With fulltext you get what you see and as there is no abstraction therefore I was more confident with the results than that of the other two.*"

Potential usages In the subjects' answers on if they would use any of the three interfaces, all indicated that they would. Context Stamp received the most favorable responses, followed by Word Cloud. Four out of thirteen participants indicated that they would use Context Stamp, four other people thought they would use Context Stamp or Word Cloud. Two participant indicated that they would use Context Stamp in parallel with the Fulltext interface, while one person would use Context Stamp or Fulltext, the rest found all three interfaces useful. Below are some of domains of interest that the subjects mentioned:

- "*I'd really like to use the Context Stamp on my collection of scholarly papers – and it would be really interesting if a digital library like ACM would use that to show how the research contexts are changing.*"
- "*Both the Word Cloud and the Context Stamp would be useful in my line of work. As a business analyst I often have to examine several documents and reports provided by our clients. In general I am either looking for something specific or I need to gain a high level understanding of the documents contents.* "
- "*I would use preferably Word Clouds or Context Stamps in domains [such as]: Gender studies [to see] in what contexts women were mentioned in formal speeches/documents/literary works over years; Religious studies how different concepts were used in different religious texts - at different times, texts of different authors, texts of different religions etc.*"
- "*I think both Context Stamps and Word Cloud would be useful tools while referring to data in the political domain (such as before election and after elections).* "
- "*I would use the Context Stamp interface for sentimental analysis, in special if I could compare two words.*"

Usability The usability ratings are shown in Figure 7.9.

We ran Friedman tests to validate the null hypotheses that there were no statistically significant differences between the three interfaces with respect to the subjective ratings on each of the attributes. The results are shown in Table 7.4.

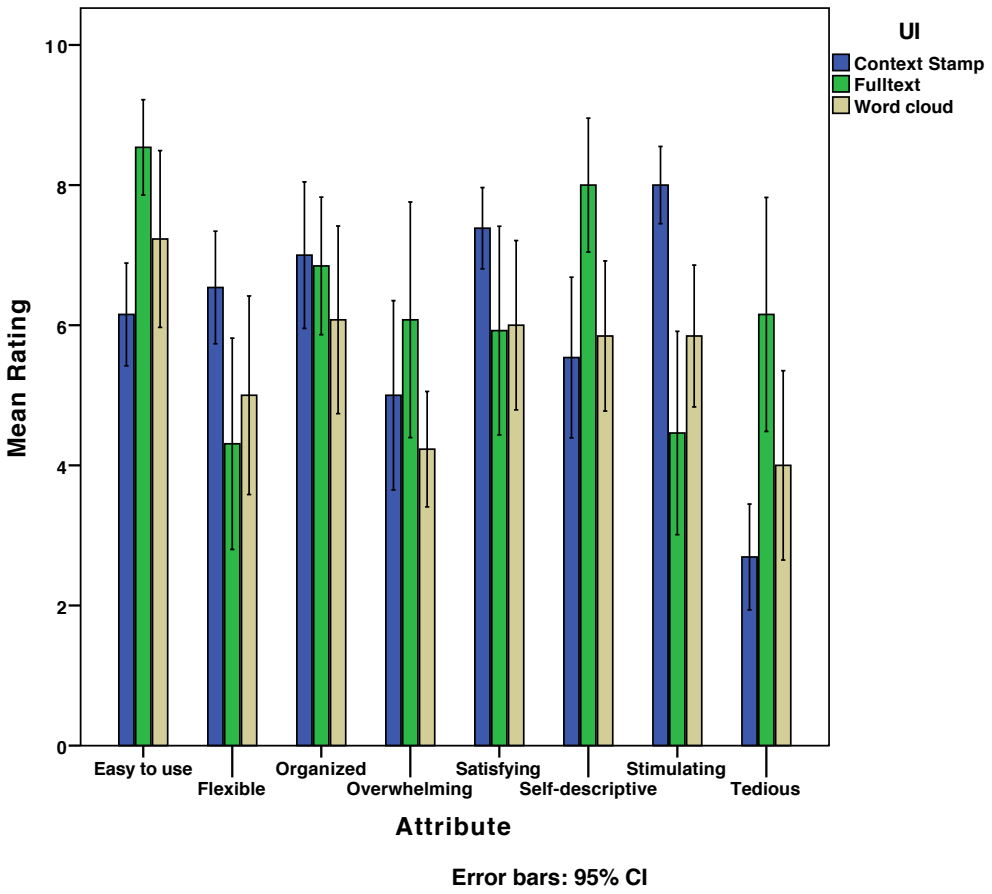


Figure 7.9: Mean Usability Ratings on the Three Interfaces

Attribute	**Test statistic & Significance**	**Pairwise Comparisons with Statistically Significant Differences**
Easy to use	$\chi^2(2) = 14.596, p = .001$	Context Stamp and Fulltext $p < .017$
Flexible	$\chi^2(2) = 10.800, p < .05$	Context Stamp and Word Cloud $p = .017$
Organized	$p > .05$	
Overwhelming	$p > .05$	
Satisfying	$p > .05$	
Self-descriptive	$\chi^2(2) = 11.021, p < .05$	Context Stamp and Fulltext $p < .017$, Word Cloud and Fulltext $p = .017$
Stimulating	$\chi^2(2) = 19.878, p < .001$	Context Stamp and Fulltext $p < .017$, Context Stamp and Word Cloud $p < .017$
Tedious	$\chi^2(2) = 12.682, p < .05$	Context Stamp and Fulltext $p < .017$

Table 7.4: Usability Ratings Statistical Tests Results

7.4.3 Discussion

Though it did not come as a surprise, we learned first-hand from this user study how challenging it was to find people who were willing to participate. Apart from those who eventually gave us their feedback, six other people dropped out of the evaluation. One person, despite having an expressed interest in data visualization, wrote back to us explicitly what she thought of the study: "*I found the questionnaire a bit long... tick the boxes is grand [OK] for an evaluation but anything more is just too much.*" Although we requested the subjects to play around with the visualizations for one to three weeks at their convenience, we did not require any minimal amount of time to be spent on them. The study might have created an impression that it required a lot more time and effort than some could afford to contribute. On the other extreme, a different subject, who was a software engineer, was so interested in the visualization prototype that she showed it to her manager and incorporated his feedback into her response. Overall, the subjects took from one to five weeks to give us feedback, and we were able to collect all the responses in almost two months after the study first started. We were able to gather a certain amount of useful, and sometimes in-depth, feedback.

The number of findings reported in Section 7.4.2 was modest, as participants only gave us some examples of what they could find. It was not a task with an intensive importance for them to exhaustively find everything of their interest. Nevertheless, it was encouraging for us to see that the participants were able to explore the given text collection. They had many serendipitous encounters on various subject matters, most often with Context Stamp, sometimes in combination with other interfaces, and even with the timeline alone in some cases.

Based on the feedback, we believe that showing a decomposition of topics within term-bearing paragraphs has an advantage over frequency-based approaches that are popular in existing works. By utilizing the assignments of words into topics in a statistical model and only using their representative words to show their proportions within a TreeMaps, we were able to abstract away words belonged to the same topical dimensions. Widely used frequency-based approaches, such as the Word Cloud interface in the study, retain all non-stopwords,

and hence result in a more cluttered display of text. This aspect can best be summarized by one of the comments: "*Word-cloud is useful for quick overview, but can get a bit messy and cluttered. Context stamps are more organised and good for getting information quickly.*"

In addition, the layout of two display panels in parallel made it easy for users to make direct comparisons of a term's usage contexts and hence was well received by the subjects: "*I liked that you could view 2 years side by side to compare.*", "*Help me determine whether the usage in the multiple paragraphs is *similar* or different. Context stamps of paragraph vs. whole document get at this.*" Interestingly, during post-hoc interviews, some subjects suggested that they would be interested in using the Context Stamp visualization to compare usage contexts not just by the different topics, but also by the sentiments within speeches at different times or by different speakers, e.g.: " *Integrate indications of emotional content – how objective vs. how much rhetoric/emotionally laden language is there?*".

The most important feedback we had was with respect to the subjects' confidence in the results presented as usage contexts. It was obvious that the Fulltext presentation was most trusted because it just simply presented the text of term-bearing paragraphs in full. For the Word Cloud interface, some participants noted that, even though some words were made more prominent because of their high frequencies, they did not necessarily represent the main points of the text well. In the case of Context Stamp, while in many cases the subjects felt that the splitting of the text into groups of topically related words made sense and understood why they were related, in other cases they did not get that, as well as disagreed with the chosen representative terms for some topics. This feedback was not surprising for us, considering the various factors that pose influences on direct human interpretations of a topic model's internal structure: (1) domain knowledge, (2) subjectivity, (3) model selection and (4) visualization of uncertainty.

With respect to the domain knowledge, despite being aware of current events and recent past history, none of the participants had full knowledge of the wide ranging matters discussed within this data set. Therefore, it is natural that some of the relationships between words were not instantly obvious to them. For instance, a subject pointed out that it was unclear, from the 2011 speech for the seed term "*research*", why one topic included terms such as "*beat*", "*sputnik*", and "*innovation*". The subject initially thought they were not at

all related, but a few minutes later realized that they actually were. Therefore, users who are unfamiliar with the data set would likely need to dig into the text to recognize or understand the contexts better.

The above example is also related to the second factor, i.e. subjectivity. On the one hand, the topic model inference process is unsupervised and typically for a collection of documents, there is no gold standard in terms of splitting the text into groups meaningfully related words. On the other hand, when the subjects explored the text collection based on the assignments of words to topics, their perception of how semantically coherent words within topics were might vary from one person to another.

The third factor, model selection, is also important, considering the fact that Context Stamp was designed to visualize directly a part of a topic model's internal structure (the topic-word distributions) for human interpretations. The key challenge here is how to select the most suitable model for the data set. As mentioned in Section 7.4.1, we followed a known good practice from a prior research to obtain a model that would fare well in predictive likelihood based metrics. However, this alone does not necessarily mean that such a model is the most interpretable for humans, because in those metrics the internal representation of a topic model is not taken into account [17]. For this purpose, the *word intrusion* and *topic intrusion* tasks were proposed in [17] to help choose an interpretable topic model based on human judgments. In the *word intrusion* task, an evaluator detects which word should not belong to a topic, and in the *topic intrusion* task, which topic should not be assigned to a document. As we mainly utilized the topic-word distributions, we applied the *word intrusion* approach, to various models obtained from the settings described in Section 7.4.1, by selecting the model with the least number of intruders within a random set of selected topics for use in Context Stamp. As model selection is beyond our scope of work / expertise, we relied on this latest known approach at the time of implementation.

The fourth factor is the need to communicate uncertainty. In a topic model, words are assigned to topics, but with different probabilities. A word can appear with high probability in one topic, but low in others. In the current design of Context Stamp, this uncertainty is not accounted for, only the topic to which a word is assigned is considered in the visualization of a topical group. In addition, showing a list of terms belonging to the same topic, upon

clicking on its representative word, was not perceived to be very helpful in the case that those terms had low probabilities of belonging to that particular topic. A subject felt that they were isolated from the text and hence she needed to peruse the paragraph's fulltext to see in what way they were related. Hence, a way to visualize uncertainty would be helpful. With the current design, users have to read the text to find that out. In future work, we will look into a way to graphically depict this information more effectively. For instance, instead of having two different pop-ups displaying a paragraph's text and the terms in that paragraph which belong to the same topic separately, they might be combined into one. This can be done by highlighting terms within a paragraph that belong to the same topic with the same color, thereby preserves the context and is yet still able to communicate topical relationships. We believe that once users are made aware of this probabilistic information, they might have a better understanding of why words are assigned to thematically related groups.

Finally, some usability issues were raised from the feedback. We discuss those that were raised by a number of participants here. With respect to the focus+context timeline, some subjects noted that they did not expect the frequency scale to change over different focused intervals of time. While this design is widely used[13], change blindness could be an issue for some users, and hence some visual cues might be useful. In the term distribution view, when a fulltext paragraph is brought up, many expressed the need to be able to move to the next or previous paragraph in the text even if it does not contain the seed term, to find out more details. For Context Stamp, some subjects noted that it could be hard to read when the corresponding paragraph is short, because its size depends on the paragraph's length. In addition, as colors are assigned randomly to topics, it might not be easy to differentiate between topics with resembling colors when the number of topics is large. Furthermore, when a document is really long, the visualization of context information may not be within the visible viewport when users point to a paragraph toward the end of that document. Therefore, some sort of overlay of the visualization might be helpful.

[13]For instance: `http://yhoo.it/c6cPVk` (Last accessed date: 29 Nov 2011)

7.5 Summary

In this chapter, we have presented our effort in visual text analytics to support users in both understanding the contexts within which a term is used as well as comparing and contrasting those between two documents. Context Stamp is an innovative approach in visual concordance analysis, it couples an advanced text mining technique (statistical topic modeling) with a combination of intuitive visualization metaphors to bring about the best of both worlds. As presenting a large set of contexts in their full text form would require a lot of effort from users to make sense of the complex and dynamic underlying meanings, our approach lets users see an easily digestible visual representation of the gist of the contexts. We conducted a user study to solicit feedback on Context Stamp and other widely used approaches in the literature. The responses on Context Stamp had been encouraging with the subject being able to reach a number of interesting and surprising findings.

The feedback was also helpful for us in future work. Visualization of uncertainty (topic-word probabilities) is an interesting area for further research. Many other additions will improve the usefulness of Context Stamp, such as the ability to analyze more than one term on the same document or multiple documents to see their correlations, as in FeatureLens [32], or automatic consideration of syntactic variations of a term (e.g. "*invest*" and "*investment*"). It would also benefit users if Context Stamp allows for different levels of document aggregation, such as aggregating all quarterly business reports within a year, to support analysis tasks. Another potential research direction is to provide a timeline-based visual representation of how contexts evolve over time, similar to the TIARA system [83] but not for document collection as a whole , but only for contextual contents for a seed term. Finally, it is also worth investigating how we can integrate an ontology representing users' knowledge into the topic model inference process, in such a way that different linguistic variations for each entity and the relationships between entities can be taken into account.

Part IV

Summary

Chapter 8

Summary and Outlook

8.1 Summary

Information visualization and visual analytics are known as essential mechanisms to support users in exploring different facets of complex information spaces. In this dissertation, we focus our attention on assisting users in the task of visual exploration of text collections. Given the amount and the unstructured or weakly structured nature of textual documents that users have to deal with, developments in visualization research are beneficial in helping them gain needed insights in a timely manner.

The contributions of this thesis are as follows:

- We propose an approach to support users in exploring text collections based on their personal interests and knowledge. In this approach, entities of interest to a user are encapsulated within an ontology, and this ontology is used to drive the exploration and analysis of a document collection. The approach is implemented in a visualization prototype called IVEA, which employs multiple coordinated views to visualize various aspects of a text collection in relation to the set of entities. IVEA enables interactions such as faceted browsing to filter for a particular subset of documents from a collection, as well as intra-document navigation based on the distribution of these entities. IVEA also enables users to incrementally enrich their ontologies with new entities matching their evolving interests in the process, and thus benefiting future usages.

- We propose an approach for faceted browsing of a collection of documents that uses a multi-dimensional visualization as an alternative to the linear listing of focus items. In this visualization, visual abstraction based on a combination of a conceptual structure and the structural equivalence of documents can be simultaneously used to deal with a large number of items. Furthermore, the approach also enables visual ordering based on the importance of facet values to support prioritized, cross-facet comparisons of focus items. A user study was conducted and the results suggested that interfaces using the proposed approach can support users better in exploratory tasks and were also well-liked by the participants of the study, with the hybrid interface combining the multi-dimensional visualization with the linear listing view receiving the most favorable ratings.

- We propose an approach to improve the visual representation of TileBars-based Entities Distribution Views. The TileBars paradigm has traditionally been used to show the distribution information of query terms in full-text documents. However, when used to show the distribution of a large number of entities of interest to users within a document, it hinders users' quick comprehension due to the visual complexity problem. Our approach employs a simplified version of a matrix reordering technique, which is based on the barycenter heuristic for bigraph edge crossing minimization, to re-arrange elements of TileBars-based Entities Distribution Views. The reordered view enables users to quickly and easily identify which entities appear in the beginning, the end, or throughout a document. A user study has suggested that even though the quantitative performance scores were not statistically significantly better with the reordered views, they were well-liked by the study's subjects and they also appreciated the benefit of the reordering operation while exploring a text document.

- We propose an approach toward content abstraction for visual concordance analysis, which supports users in understanding how terms are used within a document by investigating their usage contexts. In order to reduce the necessary effort from users to make sense of the underlying complex and dynamic semantic dimensions of contexts,

we propose Context Stamp as a visual representation of the gist of a term's usage contexts. To abstract away the textual details and yet retain the core facets of a term's contexts for visualization, we blend a statistical topic modeling method with a combination of the Treemaps and Seesoft visual metaphors. We conducted a user study on Context Stamp and two other widely used approaches in visual concordance analysis, with Context Stamp being well-received and thought to be helpful in providing support to users in understanding the usage contexts of a seed term.

Overall, the essence of our research is to enable text data owners to obtain and deepen their understanding by having a conversation with the data via suitable visualizations and associated interactions. We believe that a joyous user experience is important for this conversation to produce worthwhile outcomes. Over the course of our research, the following aspects are worth noting:

Visualizations and Interactions The design of visualizations is usually prone to legibility, perceptual, and layout challenges [144]. Hence, within our work, we try to strike a balance between aesthetic and functional properties so that the proposed visualizations can be visually engaging and at the same time, easily comprehensible, and helpful for the intended tasks. All visual components containing advanced features in IVEA and Context Stamp are extensions of existing visual metaphors that most users can understand. Care is also taken to make sure that text can be read where applicable and the layout does not become unwieldy when there are a large number of items to be displayed.

Interactions are no less important than the visualizations themselves. In our work, most interactions involve just a single mouse-click or a drag-and-drop operation. Nevertheless, from user feedback we learn that more stepwise visual cues to indicate which actions are possible next would make the application prototypes more self-descriptive.

The cold-start problem This problem usually exists for knowledge-based applications such as IVEA. While experienced users such as business analysts are most likely to have a specific set of entities of interest and importance to them, others may not. Hence this might make it difficult for some users to perceive the potential usefulness, which in turn gives them

little motivation to build an ontology to jump start the exploration. Therefore, to cater for a wide range of users with different usage scenarios, in later versions of the IVEA prototype we implemented a feature to let users start with an empty ontology and provide them with a set of suggested frequent terms / phrases from the selected text collection so that they can populate the ontology. In other words, doing so makes it easier for users to "*plant a seed and watch it grow*" while exploring text documents with IVEA.

Evaluation "*One of the perennial problems in visualization research is the difficulty of evaluating designs*"[144]. In fact, we encountered a number of challenges while carrying out evaluations in our research, chief among them are recruiting suitable participants and choosing a data set that the participants can relate to. Despite these challenges, those evaluations are certainly worthwhile. They are invaluable in many ways as we can gather and learn about user feedback in detail, especially when they are given a chance to freely use an interactive tool on their own without the pressure of being watched by researchers. The feedback has proven useful for us to understand the perceived strengths and weaknesses, as well as potential areas for future research.

8.2 Outlook

Visual exploration of text collections is a vast area with a lot of room for research and development. Depending on the application domains, there are many more potential avenues for further exploration. Here we touch on some extensions that might be relevant to IVEA, Context Stamp and both.

In IVEA, instead of providing users with a set of suggested terms or phrases appearing most frequently in a text collection, it might be also useful to employ taxonomy learning methods to suggest them with a hierarchy of concepts, from which they can adjust according to their interest. Anaphora resolution techniques may also be used so that co-references of entities, e.g., "*European Central Bank*", "*ECB*", "*the bank*", "*it*" can be identified in a text and taken into consideration. However, it is important to note that anaphora resolution is a challenging computational linguistics problem, plus an imperfect solution might introduce

noise into the data used for visualization. Furthermore, sometimes it might be necessary to enable users to disambiguate some of the variations associated with entities.

In addition, in the current implementation of IVEA, the only relationship between documents and entities is the "*Document contains Entity*" relationship. If any particular domains or tasks require, other relationships might also be taken into account. For instance, assuming that we have an ontology for the scientific publication area in which various relationships between documents and concepts, claims, explanations are as follows (for brevity / clarity we do not include the namespace here):

- *Document contains Entity*
- *Document defines Claim*
- *Document explains Concept*

With this extended model, we can have a richer way of browsing these relationships (assuming that e.g., claims can be identified in the text using framework such as SALT [51]).

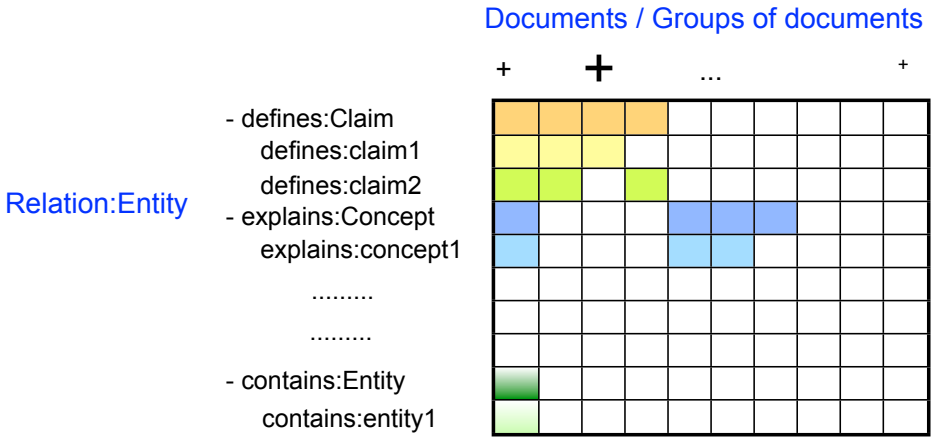


Figure 8.1: A mock-up of a visualization for faceted filtering that incorporates more semantic relations.

For instance, for faceted browsing, the matrix visualization can be adapted to indicate such relationships, as shown in Figure 8.1. The key difference from the visualization proposed in Chapter 5 is that instead of showing only the names of the classes and instances

in the hierarchy attached to the matrix, we also include the relationship names, e.g., "*defines:Claim*" to indicate which documents define any of the claim instances, and then "*defines:claim1*" to indicate which documents define the instance "*claim1*" of the class "*Claim*". This labeling scheme is applied similarly to other relations. Note that Figure 8.1 shows a mix of binary and continuous valued relationships (binary ones such as "*defines*", "*explains*", and continuous ones such as "*contains*").

In a similar fashion, we can also take these extended relationships into account when showing the distribution of entities within a document, as shown in Figure 8.2. In this Figure, users can identify which parts of a document define a claim, explain a concept, or contain certain entities.

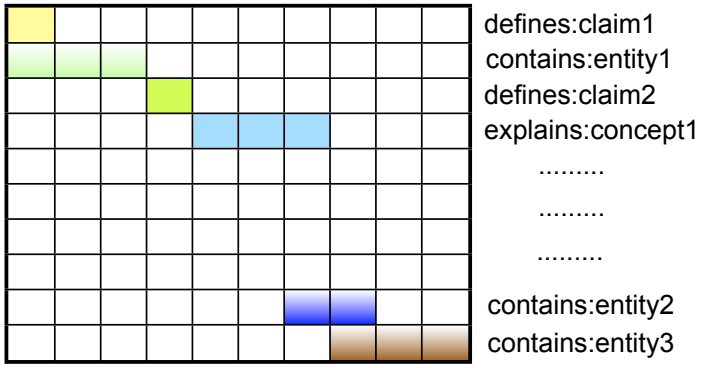


Figure 8.2: A mock-up of a visualization for an Entities Distribution View that incorporates more semantic relations.

With respect to Context Stamp, apart from some of the previously identified shortcomings that need addressing, it would be an exciting research direction to investigate how this approach can be used on the kind of short text used in social media such as tweets, status updates, etc. This kind of text poses many challenges, such as their typically short lengths and the pervasive use of colloquialisms, which make it a lot more difficult to analyze the contents.

A number of other potential extensions can be applied to both IVEA and Context Stamp. For instance, instead of dealing with only static collections of text, it will be an interesting challenge to deal with streams of text data. And when there are large streams of data coming in, distributed text processing techniques may be required to cope with the *big data* issue.

Furthermore, it might be useful to integrate personalization techniques so that the exploration process can be adaptive based on documents that users have focused on in the past. As a result, interactive tools like IVEA and Context Stamp might be able to recommend which documents might be good candidates for subsequent perusal.

In addition, at a high level, data visualization plays a powerful role as a communication enabler [144]. Viégas and Wattenberg argue that it is essential to design visualizations for communication because insights that matter will need to be communicated to others [144]. Meanwhile, "*ad hoc sharing of non-interactive versions [such as videos, screenshots] of a visualization is an unsatisfactory solution*" [144]. They believe that "*asynchronous communication of visualization-driven insights is a key aspect of communication-minded visualization*". In this regard, it would be beneficial to enable users to share the results they can obtain from interactive visualization tools with others, similar to how the ManyEyes [147] platform democratizes visualization for the masses.

Lastly, complex application domains may require solutions that enable collaborative visual text analytics, so that analysts can join force in analyzing text documents and foraging relevant information from them in a synchronous or asynchronous manner. An example of a collaborative visual analytics application is the Cambiera system [66], which employs a collaborative brushing and linking technique to enable analysts to be aware of interactions of others. This awareness might facilitate analysts in getting to the relevant documents faster and improve the quality of their findings [66]. Cambiera is used on a tabletop display in which up to four analysts can explore text collections together synchronously. We envisage both IVEA and Context Stamp can be enhanced in such a way that they can support collaboration as with Cambiera. This direction might benefit from design considerations for collaborative visual analytics discussed in [60].

8.3 Enabling Networked Knowledge

The research work reported within this dissertation was carried out at the Digital Enterprise Research Institute (DERI). The main focus of research efforts undertaken within DERI is on "*Enabling Networked Knowledge*" in order to tackle the *information overload* problem

that has been making us "*drowning in information and starving for knowledge*" [30]. Here networking of knowledge refers to a process that "*can produce a piece of knowledge whose information value is far beyond the mere sum of the individual pieces, i.e., it creates new knowledge*" [30]. The central hypothesis posed at DERI regarding networked knowledge is as follows: "*collaborative access to networked knowledge assists humans, organisations and systems with their individual as well as collective problem solving, creating solutions to problems that were previously thought insolvable, and enabling innovation and increased productivity on individual, organisational and global levels.*" [30]. In this context, research done at DERI aims at:

- develop the tools and techniques for creating, managing and exploiting networks of knowledge
- produce real-world networks of knowledge that provide maximum gains over the coming years for human, organisational and systems problem solving
- validate the hypothesis; and
- create standards supporting industrial adaptation

Among them, our work can be aligned with the first aim in that we contributed prototypical tools and techniques to help users visually explore text collections, which is a challenging task directly related to information overload problem. This exploration process takes into account users' interests and knowledge encapsulated in a user-defined ontology. As a result, the exploration can exploit the semantics contained within such an ontology, including entities of interest, their linguistic variations, and their relationships. Not only does the exploration process exploit the knowledge contained the ontology, it also facilitates managing and creating new knowledge. For instance, users can edit the entities within the ontology (e.g., change the linguistic variations list, remove unwanted entities) and add new entities matching their interest in the process. As such, there is a loop of knowledge exploitation and knowledge creation / management inherent in our work. We believe that this is of benefit to users as it allows them to keep their ontologies updated. Consequently, they can better explore a text collection when their spheres of interests are better represented. Besides, a well-tailored

personal ontology is useful not only for the text collection exploration task alone. For instance, when used within a Social Semantic Desktop environment (originally proposed in [29] and materialized in the EU IP project NEPOMUK [50]) with a PIMO ontology, such a knowledge management feature of IVEA can also benefit other PIMO-based applications. The existence of a PIMO ontology containing rich and useful information will consequently motivate users to link concepts in the ontology representing users' mental models to desktop items. This activity is certainly of significant importance to the creation, management, and exploitation of networked knowledge within Social Semantic Desktop environments, which can eventually be used to create the semantic bridges necessary for data exchange and application integration [30].

Bibliography

[1] AHLBERG, C., WILLIAMSON, C., AND SHNEIDERMAN, B. Dynamic queries for information exploration: an implementation and evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1992), CHI '92, ACM, pp. 619–626.

[2] AHMAD, W., AND KHOKHAR, A. cHawk: An Efficient Biclustering Algorithm based on Bipartite Graph Crossing Minimization. In *VLDB Workshop on Data Mining in Bioinformatics* (2007).

[3] ASSOGBA, Y., ROS, I., DIMICCO, J., AND MCKEON, M. Many Bills: Engaging Citizens through Visualizations of Congressional Legislation. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 433–442.

[4] BERNSTEIN, M. S., SUH, B., HONG, L., CHEN, J., KAIRAM, S., AND CHI, E. H. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (New York, NY, USA, 2010), UIST '10, ACM, pp. 303–312.

[5] BERTIN, J., Ed. *Graphics and graphic information-processing*. de Gruyter, Berlin, 1981.

[6] BERTIN, J. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983.

[7] BLEI, D., AND LAFFERTY, J. Topic Models. In *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

[8] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3* (March 2003), 993–1022.

[9] BOGURAEV, B., KENNEDY, C., BELLAMY, R., BRAWER, S., WONG, Y. Y., AND SWARTZ, J. Dynamic presentation of document content for rapid on-line skimming. In *AAAI Spring 1998 Symposium on Intelligent Text Summarization* (1998), pp. 118–128.

[10] BOSTOCK, M., AND HEER, J. Protovis: A graphical toolkit for visualization. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)* (2009).

[11] BYRD, D. A Scrollbar-based Visualization for Document Navigation. In *Proceedings of the fourth ACM conference on Digital libraries* (New York, NY, USA, 1999), DL '99, ACM, pp. 122–129.

[12] CARD, S. K., MACKINLAY, J. D., AND SHNEIDERMAN, B., Eds. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[13] CHABOT, C. Demystifying Visual Analytics. *IEEE Computer Graphics and Applications* (March/April 2009), 84–87.

[14] CHALMERS, M. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proceedings of the 7th conference on Visualization '96* (Los Alamitos, CA, USA, 1996), VIS '96, IEEE Computer Society Press, pp. 127–ff.

[15] CHALMERS, M., AND CHITSON, P. Bead: Explorations in information visualization. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1992), SIGIR '92, ACM, pp. 330–337.

[16] CHALMERS, M., INGRAM, R., AND PFRANGER, C. Adding Imageability Features to Information Displays. In *Proceedings of the 9th annual ACM symposium on User interface software and technology* (New York, NY, USA, 1996), UIST '96, ACM, pp. 33–39.

[17] CHANG, J., BOYD-GRABER, J., WANG, C., GERRISH, S., AND BLEI, D. M. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS* (2009).

[18] CHI, E. H. A Taxonomy of Visualization Techniques Using the Data State Reference Model. In *Proceedings of the IEEE Symposium on Information Vizualization 2000* (Washington, DC, USA, 2000), INFOVIS '00, IEEE Computer Society, pp. 69–75.

[19] CHI, E. H., AND RIEDL, J. An Operator Interaction Framework for Visualization Systems. In *Proceedings of the 1998 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1998), IEEE Computer Society, pp. 63–70.

[20] CLARKSON, E., DESAI, K., AND FOLEY, J. ResultMaps: Visualization for Search Interfaces. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 1057–1064.

[21] CLARKSON, E. C., NAVATHE, S. B., AND FOLEY, J. D. Generalized formal models for faceted user interfaces. In *JCDL '09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2009), ACM, pp. 125–134.

[22] COLLINS, C., CARPENDALE, S., AND PENN, G. DocuBurst: Visualizing Document Content using Language Structure. *Computer Graphics Forum 28*, 3 (2009), 1039–1046.

[23] COLLINS, C., VIGAS, F. B., AND WATTENBERG, M. Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2009), pp. 91–98.

[24] COOK, K., EARNSHAW, R., AND STASKO, J. Guest editors' introduction: Discovering the unexpected. *IEEE Computer Graphics and Applications 27* (2007), 15–19.

[25] CROSSNO, P., DUNLAVY, D., AND SHEAD, T. LSAView: A tool for visual exploration of latent semantic modeling. In *VAST 2009: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2009.* (2009), pp. 83–90.

[26] CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., AND TABLAN, V. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics* (2002).

[27] D'AQUIN, M., BALDASSARE, C., GRIDINOC, L., SABOU, M., ANGELETOU, S., AND MOTTA, E. Watson: Supporting Next Generation Semantic Web Applications. In *WWW/Internet conference 2007* (2007).

[28] DECAMP, P., FRID-JIMENEZ, A., GUINESS, J., AND ROY, D. Gist Icons: Seeing Meaning in Large Bodies of Literature. In *InfoVis '05: Proceedings of the IEEE Information Visualization 2005 Conference* (2005).

[29] DECKER, S., AND FRANK, M. The Social Semantic Desktop. In *Workshop on Application Design, Development and Implementation Issues in the Semantic Web at the 13th International World Wide Web Conference* (New York, NY, USA, May 2004).

[30] DECKER, S., AND HAUSWIRTH, M. Enabling networked knowledge. In *Cooperative Information Agents XII*, vol. 5180 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008, pp. 1–15.

[31] DHILLON, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. *Knowledge Discovery and Data Mining*, 3 (2001), 269–274.

[32] DON, A., ZHELEVA, E., GREGORY, M., TARKAN, S., AUVIL, L., CLEMENT, T., SHNEIDERMAN, B., AND PLAISANT, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *CIKM '07: Proceedings of the 16th ACM Conference on information and Knowledge Management* (New York, NY, USA, 2007), ACM, pp. 213–222.

[33] DÖRK, M., CARPENDALE, S., COLLINS, C., AND WILLIAMSON, C. VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1205–1212.

[34] DREDZE, M., WALLACH, H. M., PULLER, D., AND PEREIRA, F. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces* (New York, NY, USA, 2008), IUI '08, ACM, pp. 199–206.

[35] EICK, S. G. Graphically displaying text. *Journal of Computational and Graphical Statistics 3* (1994), 127–142.

[36] EICK, S. G., AND KARR, A. F. Visual scalability. *Journal of Computational and Graphical Statistics 11*, 1 (Mar 2002), 22–43.

[37] EICK, S. G., STEFFEN, J. L., AND SUMNER, JR., E. E. Seesoft-A Tool for Visualizing Line Oriented Software Statistics. *IEEE Trans. Softw. Eng. 18* (November 1992), 957–968.

[38] FEKETE, J.-D., AND DUFOURNAUD, N. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), ACM, pp. 47–55.

[39] FELLBAUM, C. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[40] FERRÉ, S. Agile Browsing of a Document Collection with Dynamic Taxonomies. In *DEXA '08: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 377–381.

[41] FIRTH, J. R. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis* (1957), 1–32.

[42] FLUIT, C. AutoFocus: Semantic Search for the Desktop. In *IV'05: Proceedings of the Ninth International Conference on Information Visualisation* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 480–487.

[43] FURNAS, G. W. Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1986), CHI '86, ACM, pp. 16–23.

[44] FURNAS, G. W., DEERWESTER, S., DUMAIS, S. T., LANDAUER, T. K., HARSHMAN, R. A., STREETER, L. A., AND LOCHBAUM, K. E. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1988), SIGIR '88, ACM, pp. 465–480.

[45] GERKEN, J., BAK, P., AND REITERER, H. Longitudinal Evaluation Methods in Human-Computer Studies and Visual Analytics . In *Metrics for the Evaluation of Visual Analytics (an InfoVIS 2007 Workshop)* (Oct 2007).

[46] GIRGENSOHN, A., SHIPMAN, F., CHEN, F., AND WILCOX, L. DocuBrowse: faceted searching, browsing, and recommendations in an enterprise context. In *IUI '10: Proceeding of the 14th international conference on Intelligent user interfaces* (New York, NY, USA, 2010), ACM, pp. 189–198.

[47] GRIDINOC, L., D'AQUIN, M., DZBOR, M., LOPEZ, V., AND MOTTA, E. D8.3 - Final Version of the Semantic Web Browser. Tech. rep., Knowledge Media Institute, The Open University, UK, 2008.

[48] GRIDINOC, L., SABOU, M., D'AQUIN, M., DZBOR, M., AND MOTTA, E. Semantic Browsing with PowerMagpie. In *The Semantic Web: Research and Applications*, vol. 5021 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2008, pp. 802–806.

[49] GRIFFITHS, T. L., AND STEYVERS, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101*, Suppl 1 (Apr. 2004), 5228–5235.

[50] GROZA, T., HANDSCHUH, S., MOELLER, K., GRIMNES, G., SAUERMANN, L., MINACK, E., MESNAGE, C., JAZAYERI, M., REIF, G., AND GUDJONSDOTTIR, R. The NEPOMUK Project – On the way to the Social Semantic Desktop. In *Proc. of the 3rd Int'l Conf. on Semantic Technologies (I-SEMANTICS 2007), Graz, Austria* (2007).

[51] GROZA, T., MÖLLER, K., HANDSCHUH, S., TRIF, D., AND DECKER, S. Salt: weaving the claim web. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference* (Berlin, Heidelberg, 2007), ISWC'07/ASWC'07, Springer-Verlag, pp. 197–210.

[52] HAMMING, R. W. *Visual Explanations: Images and Quantities, Evidence and Narrative*. McGraw-Hill, 1973.

[53] HAVRE, S., HETZLER, E., WHITNEY, P., AND NOWELL, L. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics 8* (January 2002), 9–20.

[54] HEALEY, C. G., BOOTH, K. S., AND ENNS, J. T. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction 3* (June 1996), 107–135.

[55] HEARST, M. A. Multi-Paragraph Segmentation of Expository Text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 1994), ACL '94, Association for Computational Linguistics, pp. 9–16.

[56] HEARST, M. A. TileBars: visualization of term distribution information in full text information access. In *CHI '95: Proceedings of the SIGCHI Conference on Human factors in computing systems* (New York, NY, USA, 1995), ACM, pp. 59–66.

[57] HEARST, M. A. Clustering versus faceted categories for information exploration. *Communications of the ACM 49*, 4 (2006), 59–61.

[58] HEARST, M. A. UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In *HCIR '08: Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval* (October 2008).

[59] HEARST, M. A. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 2009.

[60] HEER, J., AND AGRAWALA, M. Design Considerations for Collaborative Visual Analytics. *Information Visualization* (2008), 49–62.

[61] HEER, J., CARD, S. K., AND LANDAY, J. A. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceedings of the SIGCHI Conference on Human factors in computing systems* (New York, NY, USA, 2005), ACM Press, pp. 421–430.

[62] HILDEBRAND, M., VAN OSSENBRUGGEN, J., AND HARDMAN, L. /facet: A Browser for Heterogeneous Semantic Web Repositories. In *Proceedings of the 5th international conference on The Semantic Web* (Berlin, Heidelberg, 2006), ISWC'06, Springer-Verlag, pp. 272–285.

[63] HUANG, S., WARD, M. O., AND RUNDENSTEINER, E. A. Exploration of dimensionality reduction for text visualization. In *Proceedings of the Coordinated and Multiple Views in Exploratory Visualization* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 63–74.

[64] HUYNH, D., KARGER, D., AND MILLER, R. Exhibit: Lightweight Structured Data Publishing. In *WWW '07: Proceedings of the 16th International World Wide Web Conference* (Banff, Alberta, Canada, 2007), ACM.

[65] INSELBERG, A., AND DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90: Proceedings of the 1st Conference on Visualization '90* (Los Alamitos, CA, USA, 1990), IEEE Computer Society Press, pp. 361–378.

[66] ISENBERG, P., AND FISHER, D. Collaborative Brushing and Linking for Co-located Visual Analytics of Document Collections. *Computer Graphics Forum 28*, 3 (June 2009), 1031–1038.

[67] JOLLIFFE, I. *Principal Component Analysis*. Springer Verlag, 1986.

[68] KÄKI, M., AND AULA, A. Controlling the complexity in comparing search user interfaces via user studies. *Information Processing and Management 44*, 1 (2008), 82–91.

[69] KASKI, S., HONKELA, T., LAGUS, K., AND KOHONEN, T. WEBSOM - Self-organizing maps of document collections. *Neurocomputing 21*, 1-3 (Nov. 1998), 101–117.

[70] KEIM, D., ANDRIENKO, G., FEKETE, J.-D., GÖRG, C., KOHLHAMMER, J., AND MELANON, G. Visual analytics: Definition, process, and challenges. In *Information Visualization*, A. Kerren, J. Stasko, J.-D. Fekete, and C. North, Eds., vol. 4950 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2008, pp. 154–175.

[71] KEIM, D. A., BAK, P., BERTINI, E., OELKE, D., SPRETKE, D., AND ZIEGLER, H. Advanced visual analytics interfaces. In *Proceedings of the International Conference on Advanced Visual Interfaces* (New York, NY, USA, 2010), AVI '10, ACM, pp. 3–10.

[72] KEIM, D. A., KOHLHAMMER, J., ELLIS, G., AND MANSMANN, F., Eds. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, November 2010.

[73] KEIM, D. A., MANSMANN, F., AND THOMAS, J. Visual analytics: how much visualization and how much analytics? *SIGKDD Explorations Newsletter 11* (May 2010), 5–8.

[74] KEIM, D. A., AND OELKE, D. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *VAST 2007: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2007* (2007), pp. 115–122.

[75] KOHONEN, T. *Self-organization and associative memory: 3rd edition*. Springer-Verlag New York, Inc., New York, NY, USA, 1989.

[76] KOHONEN, T., SCHROEDER, M. R., AND HUANG, T. S., Eds. *Self-Organizing Maps*, 3rd ed. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[77] KORMAN, R. INVESTING IT; Mining for Nuggets Of Financial Data. `http://www.nytimes.com/1998/06/21/business/investing-it-mining-for-nuggets-of-financial-data.html`, 1998.

[78] KRUSKAL, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (Mar. 1964), 1–27.

[79] KULES, B., AND CAPRA, R. Creating exploratory tasks for a faceted search interface. In *HCIR '08: Proceedings of the Second Workshop on Human-Computer Interaction* (2008).

[80] KULES, B., CAPRA, R., BANTA, M., AND SIERRA, T. What do exploratory searchers look at in a faceted search interface? In *JCDL '09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2009), ACM, pp. 313–322.

[81] LEE, B., SMITH, G., ROBERTSON, G. G., CZERWINSKI, M., AND TAN, D. S. FacetLens: exposing trends and relationships to support sensemaking within faceted datasets. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), ACM, pp. 1293–1302.

[82] LIN, X. Map displays for information retrieval. *Journal of the American Society for Information Science 48* (January 1997), 40–54.

[83] LIU, S., ZHOU, M. X., PAN, S., QIAN, W., CAI, W., AND LIAN, X. Interactive, topic-based visual text summarization and analysis. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management* (New York, NY, USA, 2009), ACM, pp. 543–552.

[84] LORRAIN, F., AND WHITE, H. C. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology 1* (1971), 49–80.

[85] MACKINLAY, J. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics 5* (April 1986), 110–141.

[86] MANN, T. M. *Visualization of search results from the World Wide Web*. PhD thesis, University of Konstanz, 2002.

[87] MANNING, C. D., AND SCHUETZE, H. *Foundations of Statistical Natural Language Processing*, 1 ed. The MIT Press, June 1999.

[88] MCCALLUM, A. K. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[89] MCCORMICK, B., DEFANTI, T., AND BROWN, M. Visualization in Scientific Computing. *Computer Graphics 21* (November 1987).

[90] MOTHE, J., CHRISMENT, C., DOUSSET, B., AND ALAUX, J. DocCube: multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology 54*, 7 (2003), 650–659.

[91] NEUMANN, P. Focus+Context Visualization of Relations in Hierarchical Data. Diploma thesis, Otto-von-Guericke University Magdeburg, 2004.

[92] NORMAN, D. A. *Things that make us smart : defending human attributes in the age of the machine*. Addison-Wesley Pub. Co., 1993.

[93] NORTH, C. Toward Measuring Visualization Insight. *IEEE Computer Graphics and Applications 26*, 3 (2006), 6–9.

[94] OELKE, D., BAK, P., KEIM, D. A., LAST, M., AND DANON, G. Visual Evaluation of Text Features for Document Summarization and Analysis. In *In IEEE Symposium on Visual Analytics Science and Technology (VAST* (2008), pp. 75–82.

[95] OLSEN, K. A., KORFHAGE, R. R., SOCHATS, K. M., SPRING, M. B., AND WILLIAMS, J. G. Visualization of a document collection: the VIBE system. *Information Processing and Management 29*, 1 (1993), 69–81.

[96] OREN, E., DELBRU, R., AND DECKER, S. Extending faceted navigation for RDF data. In *Proceedings of the 5th international conference on The Semantic Web* (Berlin, Heidelberg, 2006), ISWC'06, Springer-Verlag, pp. 559–572.

[97] PALEY, W. B. TextArc: Showing Word Frequency and Distribution in Text. In *Poster Proc of IEEE InfoVis* (2002).

[98] PAULOVICH, F. V., OLIVEIRA, M. C. F., AND MINGHIM, R. The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization. In *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 27–36.

[99] PIROLLI, P., AND CARD, S. Information foraging. *Psychological Review 106. 4* (1999), 634–675.

[100] PLAISANT, C., BRUNS, T., SHNEIDERMAN, B., AND DOAN, K. Query previews in networked information systems: the case of EOSDIS. In *CHI '97: Extended abstracts on Human factors in computing systems* (New York, NY, USA, 1997), ACM, pp. 202–203.

[101] PLAISANT, C., FEKETE, J.-D., AND GRINSTEIN, G. Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics 14*, 1 (2008), 120–134.

[102] RANGANATHAN, S. *Colon classification, Basic Classification*. Asia Publishing House, 1933.

[103] RAO, R., AND CARD, S. K. The Table Lens: Merging Graphical and Symbolic. Representations in an Interactive Focus+Context. Visualization for Tabular Information. In *CHI '94: Proceedings of the SIGCHI Conference on Human factors in computing systems* (New York, NY, USA, 1994), ACM, pp. 318–322.

[104] REITERER, H., MUSSLER, G., MANN, T., AND HANDSCHUH, S. INSYDER - An Information Assistant for Business Intelligence. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR 2000 Conference on Research and Development in Information Retrieval* (2000), ACM press, pp. 112–119.

[105] RHYNE, T.-M., TORY, M., MUNZNER, T., WARD, M. O., JOHNSON, C., AND LAIDLAW, D. H. Information and Scientific Visualization: Separate but Equal or Happy Together at Last. In *IEEE Visualization* (2003), IEEE Computer Society, pp. 619–621.

[106] RICHE, N. H., LEE, B., AND CHEVALIER, F. iChase: supporting exploration and awareness of editing activities on Wikipedia. In *Proceedings of the International Conference on Advanced Visual Interfaces* (New York, NY, USA, 2010), AVI '10, ACM, pp. 59–66.

[107] RISCH, J., KAO, A., POTEET, S., AND WU, Y.-J. J. Text Visualization for Visual Text Analytics. In *Visual Data Mining*, vol. 4404 of *Lecture Notes in Computer Science*. Springer, 2008, pp. 154–171.

[108] RISCH, J. S., REX, D. B., DOWSON, S. T., WALTERS, T. B., MAY, R. A., AND MOON, B. D. Readings in information visualization. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, ch. The STARLIGHT information visualization system, pp. 551–560.

[109] ROBERTS, J. C. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *CMV '07: Proc. of the 5th Int'l Conf. on Coordinated and Multiple Views in Exploratory Visualization* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 61–71.

[110] ROSE, S., BUTNER, S., COWLEY, W., GREGORY, M., AND WALKER, J. Describing Story Evolution From Dynamic Information Streams. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (October 2009), pp. 99 –106.

[111] SALTON, G., Ed. *Automatic Text Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1988.

[112] SAMMON, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers 18* (May 1969), 401–409.

[113] SAUERMANN, L., VAN ELST, L., AND DENGEL, A. PIMO - a Framework for Representing Personal Information Models. In *Proceedings of I-Semantics' 07* (2007), T. Pellegrini and S. Schaffert, Eds., JUCS, pp. 270–277.

[114] SCHRAEFEL, M. C., SMITH, D. A., OWENS, A., RUSSELL, A., HARRIS, C., AND WILSON, M. The Evolving mSpace Platform: Leveraging the Semantic. Web on the Trail of the Memex. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia* (New York, NY, USA, 2005), HYPERTEXT '05, ACM, pp. 174–183.

[115] SCHWARTZ, M., HASH, C., AND LIEBROCK, L. M. Term Distribution Visualizations with Focus+Context. In *Proceedings of the 2009 ACM symposium on Applied Computing* (New York, NY, USA, 2009), SAC '09, ACM, pp. 1792–1799.

[116] SEELING, C., AND BECKS, A. Analysing Associations of Textual and Relational Data with a Multiple Views System. In *CMV '04: Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 61–70.

[117] SHAW, C. D., KUKLA, J. M., SOBOROFF, I., EBERT, D. S., NICHOLAS, C. K., ZWA, A., MILLER, E. L., AND ROBERTS, D. A. Interactive Volumetric Information Visualization for Document Corpus Management. *Internationl Journal on Digital Libraries 2*, 2-3 (1999), 144–156.

[118] SHLENS, J. A Tutorial on Principal Component Analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies* (2005).

[119] SHNEIDERMAN, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics 11* (January 1992), 92–99.

[120] SHNEIDERMAN, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Visual Languages* (1996), pp. 336–343.

[121] SHNEIDERMAN, B., AND PLAISANT, C. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization* (New York, NY, USA, 2006), BELIV '06, ACM, pp. 1–7.

[122] SIIRTOLA, H., AND MÄKINEN, E. Constructing and reconstructing the reorderable matrix. *Information Visualization 4*, 1 (2005), 32–48.

[123] SMITH, G., CZERWINSKI, M., MEYERS, B., ROBBINS, D., ROBERTSON, G., AND TAN, D. S. FacetMap: A Scalable Search and Browse Visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 797–804.

[124] SPENKE, M., BEILKEN, C., AND BERLAGE, T. FOCUS: the interactive table for product comparison and selection. In *UIST '96: Proceedings of the 9th annual ACM symposium on User interface software and technology* (New York, NY, USA, 1996), ACM, pp. 41–50.

[125] SPOERRI, A. InfoCrystal: a visual tool for information retrieval & management. In *Proceedings of the second international conference on Information and knowledge management* (New York, NY, USA, 1993), CIKM '93, ACM, pp. 11–20.

[126] STASKO, J., GÖRG, C., AND LIU, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization 7* (April 2008), 118–132.

[127] STASKO, J., AND ZHANG, E. Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In *Proceedings of the IEEE Symposium on Information Visualization 2000* (Washington, DC, USA, 2000), INFOVIS '00, IEEE Computer Society, pp. 57–65.

[128] STEFANER, M., FÉRRÉ, S., PERUGINI, S., KOREN, J., AND ZHANG, Y. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*, vol. 25 of *The*

Information Retrieval Series. Springer, 2009, ch. 4 - User Interface Design, pp. 75–112.

[129] STEFANER, M., AND MULLER, B. Elastic lists for facet browsers. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications* (Washington, DC, USA, 2007), DEXA '07, IEEE Computer Society, pp. 217–221.

[130] STEFANER, M., URBAN, T., AND SEEFELDER, M. Elastic Lists for Facet Browsing and Resource Analysis in the Enterprise. In *Proceedings of the International Workshop on Database and Expert Systems Applications* (Los Alamitos, CA, USA, 2008), IEEE Computer Society, pp. 397–401.

[131] STOFFEL, A., SPRETKE, D., KINNEMANN, H., AND KEIM, D. A. Enhancing Document Structure Analysis using Visual Analytics. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (New York, NY, USA, 2010), SAC '10, ACM, pp. 8–12.

[132] STROBELT, H., OELKE, D., ROHRDANTZ, C., STOFFEL, A., KEIM, D. A., AND DEUSSEN, O. Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics 15* (November 2009), 1145–1152.

[133] SUGIYAMA, K., TAGAWA, S., AND TODA, M. Methods for Visual Understanding of Hierarchical System Structures. *IEEE Transactions on Systems, Man, and Cybernetics 11*, 2 (1981), 109–125.

[134] SUH, B., CHI, E. H., KITTUR, A., AND PENDLETON, B. A. Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashboard. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1037–1040.

[135] THAI, V., DAVIS, B., O'RIAIN, S., O'SULLIVAN, D., AND HANDSCHUH, S. Semantically Enhanced Passage Retrieval for Business Analysis Activity. In *ECIS 2008:*

Proceedings of the 16th European Conference on Information Systems (Galway, Ireland, 2008).

[136] THAI, V., AND HANDSCHUH, S. Enhanced Navigation and Focus on TileBars with Barycenter Heuristic-based Reordering. In *AVI '10: Proceedings of the International Conference on Advanced Visual Interfaces* (New York, NY, USA, 2010), ACM, pp. 349–352.

[137] THAI, V., AND HANDSCHUH, S. Context Stamp: A Topic-based Content Abstraction for Visual Concordance Analysis. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (New York, NY, USA, 2011), CHI EA '11, ACM, pp. 2269–2274.

[138] THAI, V., ROUILLE, P.-Y., AND HANDSCHUH, S. Visual abstraction and ordering in faceted browsing of text collections. *ACM Transactions on Intelligent Systems and Technology (ACM TIST) 3*, 2 (Feb. 2012), 21:1–21:24.

[139] THOMAS, J. J., AND COOK, K. A., Eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, Los Alametos, CA, 2005.

[140] TUFTE, E. R. *Visual Explanations: Images and Quantities, Evidence and Narrative*, first ed. Graphics Press, February 1997.

[141] TWEEDIE, L. Characterizing interactive externalizations. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1997), ACM Press, pp. 375–382.

[142] VAN HAM, F., WATTENBERG, M., AND VIÉGAS, F. B. Mapping Text with Phrase Nets. *IEEE Transactions on Visualization and Computer Graphics 15* (November 2009), 1169–1176.

[143] VIÉGAS, F. B., GOLDER, S., AND DONATH, J. Visualizing Email Content: Portraying Relationships from Conversational Histories. In *Proceedings of the SIGCHI*

conference on Human Factors in computing systems (New York, NY, USA, 2006), CHI '06, ACM, pp. 979–988.

[144] VIÉGAS, F. B., AND WATTENBERG, M. Communication-minded visualization: A call to action. *IBM Systems Journal 45*, 4 (Sept. 2006).

[145] VIÉGAS, F. B., WATTENBERG, M., AND DAVE, K. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2004), CHI '04, ACM, pp. 575–582.

[146] VIÉGAS, F. B., WATTENBERG, M., AND FEINBERG, J. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 1137–1144.

[147] VIÉGAS, F. B., WATTENBERG, M., VAN HAM, F., KRISS, J., AND MCKEON, M. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics 13* (November 2007), 1121–1128.

[148] VUILLEMOT, R., CLEMENT, T., PLAISANT, C., AND KUMAR, A. What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections. In *VAST 2009: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2009.* (2009), pp. 107–114.

[149] WALLACH, H. M., MIMNO, D., AND MCCALLUM, A. Rethinking LDA: Why priors matter. In *NIPS '09: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems* (2009).

[150] WARE, C. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[151] WATTENBERG, M., AND VIÉGAS, F. B. The Word Tree, an Interactive Visual Concordance. *IEEE Trans. Visualization & Computer Graphics 14* (2008), 1221–1228.

[152] WATTENBERG, M., VIÉGAS, F. B., AND HOLLENBACH, K. Visualizing Activity on Wikipedia with Chromograms. In *Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction - Volume Part II* (Berlin, Heidelberg, 2007), INTERACT'07, Springer-Verlag, pp. 272–287.

[153] WEI, F., LIU, S., SONG, Y., PAN, S., ZHOU, M. X., QIAN, W., SHI, L., TAN, L., AND ZHANG, Q. TIARA: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2010), KDD '10, ACM, pp. 153–162.

[154] WILSON, A. T., AND CHEW, P. A. Term weighting schemes for Latent Dirichlet Allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), HLT '10, Association for Computational Linguistics, pp. 465–473.

[155] WISE, J. A. The Ecological Approach to Text Visualization. *Journal of the American Society for Information Science - Speical issue on integrating mutiple overlapping metadata standards 50* (November 1999), 1224–1233.

[156] WISE, J. A., THOMAS, J. J., PENNOCK, K., LANTRIP, D., POTTIER, M., SCHUR, A., AND CROW, V. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1995), IEEE Computer Society, p. 51.

[157] WURMAN, R. S. *Information Anxiety 2 (Hayden/Que)*, 2nd ed. Que, Dec. 2000.

[158] YAO, L., MIMNO, D., AND MCCALLUM, A. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2009), KDD '09, ACM, pp. 937–946.

[159] YEE, K.-P., SWEARINGEN, K., LI, K., AND HEARST, M. Faceted metadata for image search and browsing. In *CHI '03: Proc. of the SIGCHI Conf. on Human factors in computing systems* (New York, NY, USA, 2003), ACM, pp. 401–408.

[160] ZHANG, J., AND MARCHIONINI, G. Evaluation and evolution of a browse and search interface: Relation Browser++. In *Proceedings of the 2005 national conference on Digital government research* (2005), dg.o '05, Digital Government Society of North America, pp. 179–188.

[161] ZHU, W., AND CHEN, C. Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics 31*, 3 (2007), 338–349.