Aspect Grid: A Visualization for Iteratively Refining Aspect-Based Queries on Document Collections

Florian Haag, Qi Han, Markus John, Thomas Ertl

Institute for Visualization and Interactive Systems, University of Stuttgart
Universitätsstraße 38 · 70569 Stuttgart · Germany
{ florian.haag, qi.han, markus.john, thomas.ertl }@vis.uni-stuttgart.de

Abstract: Search results comprising of unstructured documents are traditionally displayed as plain lists, whereas the formulation of complex queries can be aided by comprehensive visualizations that highlight relevant attributes of the search results. We propose a visualization concept that incorporates a sentiment analysis of aspects mentioned in examined documents. Our concept provides a means of iteratively refining the query based upon preliminary search results. After an abstract description of the concept, we elaborate on its usage in a practical example of filtering customer reviews in order to choose a restaurant based on user-defined preferences.

1 Introduction

Traditionally, search engines expect an enumeration of features being looked for—such as keywords that must appear in a document—and results are displayed as a purely sequential list of references to the found documents. This generic way of searching is suitable for the web with its diverse kinds of resources. However, users who look for some particular information often require more specific queries and start using advanced operators [AKG10]. In addition, they are interested only in particular types of web resources—for instance, customers looking for a hotel offer wants to find only web resources that contain information on hotel rates, they do not want to find artistic images, phone book entries or recipes. Accordingly, search engines can offer means for defining complex queries and improve the overview of search results by taking advantage of the special traits of a given resource kind [Do08].

Often, such querying systems rely on structured information available along with the search results. This is particularly true in query systems for the Semantic Web [St12], but can also be observed with other systems, such as image search engines that examine metadata [Ye03]. While users can examine the search results to manually refine their query, a direct means of extracting further query restrictions from an intermediary result set and incorporating them into the query is often missing. This is complicated by the fact that documents of the same type, but from different data sources, may feature incompatible schemas of structured information, if any, leaving search engines with the unstructured textual part of the resource as the only guaranteed common feature. We propose Aspect

Grid, a novel approach that comprises of a natural language processing component extracting relevant information from documents, combined with a user interface that provides an overview of the active query criteria, as well as options to directly refine the query based upon the extracted information.

2 Related Work

Results are displayed in a way specific to their type by domain-specific systems, such as search engines specialized on summarizing news articles [Ra05], as well as by domain-aware multi-purpose systems that apply different output rules to a variety of search result types [Bo10]. There have been attempts to structure these results, by creating clusters of similar subsets [Hi11], by providing an overview of various aspects of each search result [Iw12], by embedding search results into other visualizations such as maps or timelines, and by combining several types of search results in a compound view [Bo11]. Moreover, the search results can be organized into nodes of a topic-map [MP13] or be directly displayed as a graph [RTB11].

The expressiveness of the queries for which search results are retrieved can be increased by using faceted approaches that apply individual filter criteria to different aspects of the dataset. As large amounts of data are available in the form of unstructured text documents, the language processing community has been researching methods for the automatical extraction of structured information from unstructured data for a long time [MRS08]. Among these, sentiment analysis has received much attention recently [Li12]. Its goals are the detection and extraction of subjective information from documents, especially reviews, using supervised or unsupervised methods [PL08, BE10]. Systems that summarize and visualize sentiment information from a set of documents have been successfully implemented [Ma11].

While there are various visualization concepts that provide an overview of the facets of search results, as well as of the information extracted from unstructured documents, few of them are combined with a visual representation of the active query. Again, techniques from the Semantic Web tend to support that kind of view by integrating the search results into a query graph [St12, He09]. Many others, however, display results merely in a separate area that has no direct graphical connection to the visualization of the query [Ru08]. Only few specialized systems incorporate a direct means of iteratively refining the query based on the current result set [Ko09], or deriving a more precise query from intermediary results [Ha06].

In order to allow for a high degree of expressivity and intuitively show the currently active query, search parameters and their mutual connections—such as the logical operators *and* and *or*—, the query expression can be visualized. A variety of approaches for query visualization has been proposed, based upon a multitude of paradigms such as Venn diagrams [Jo98], flowcharts [Sh94, Ha06], graphs [Ru08, St12], or custom visual elements [Sp93]. While they successfully indicate the active search parameters and are to some extent tightly integrated with the search results, they usually do not provide a direct

feedback channel for query refinement based on the results; the query has to be updated manually by the user after looking at the results.

3 The Aspect Grid

In this work, we present the Aspect Grid approach—a visualization technique for filtering collections of text documents that integrates search parameters and results with the possibility of refining the query based upon text excerpts from search results. Two sketches of different states of the visualization are presented in Figure 1. Based upon a collection

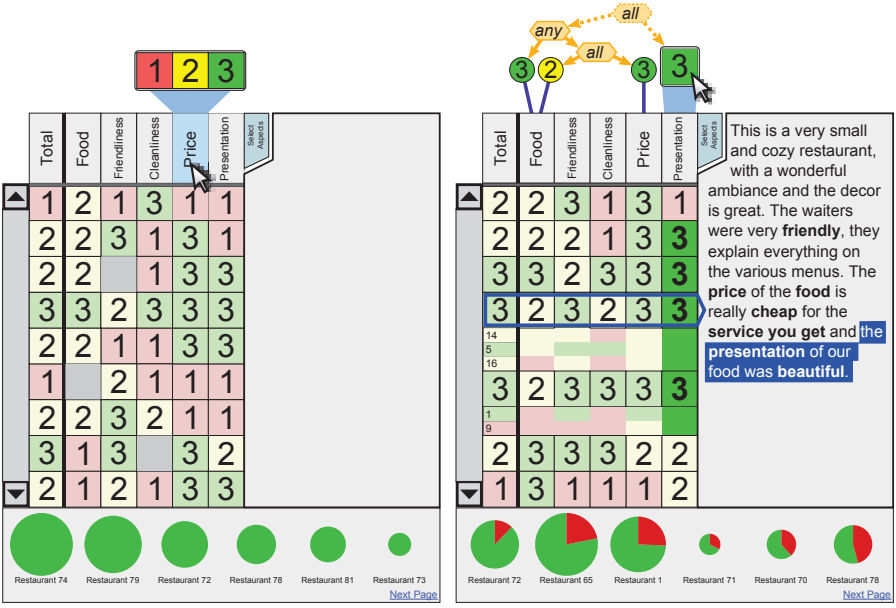

Figure 1: Two depictions of the Aspect Grid visualization being used for filtering restaurant reviews: Left: All reviews are displayed as no filters have been set up yet. The user is about to select a restriction for the aspect *Price*. Right: Several filter criteria have been set up and combined with *any* and *all* operators. The user has opened one review in the preview area and selected a statement in that review that positively mentions the *Presentation* aspect. An according filter is suggested, and as the mouse cursor hovers over it, any rows with suitable values are highlighted by intensifying the background color. The suggested filter is automatically linked with the *all* operator.

of text documents from a given domain, our approach first determines important aspects in the domain. While this is currently done by using a domain-specific prepared list of aspects, in the future we want to use more advanced techniques to automatically extract the important aspects [BE10]. Once knowing which aspects are relevant for the current domain, our system scans each document for mentions of any of the aspects. It then further calculates the sentiment scores expressed towards the detected aspects using a sentiment

lexicon-based method, as described by Liu [Li12]. We choose this method because it works without a domain-specific training dataset. We assign colors to sentiment scores to make it easier for users to navigate the results. Our sentiment color scores are one for negative (red), two for neutral (yellow), three for positive (green) and for no finding we use a empty fields (gray). When a domain-specific dataset annotated with aspects and sentiment scores is available, the accuracy of our approach can be further improved using supervised method.

Both the aspects and the recognized sentiments are reflected in the visualization. The central visual part of the Aspect Grid approach is a grid whose columns correspond with the various aspects for which statements were found in the texts. At first, each document that belongs to the general category of texts being examined occupies one row in the grid. Documents that were excluded by the filter criteria are clustered based on their aspects and reduced in size. The query to the underlying search engine is expressed graphically by depicting the restrictions for each aspect right above the respective column. In order to define different combinations (*all*, *any*) of aspect restrictions, these restrictions are visually linked among each other.

The third part of the visualization is an area that can be used to display the contents of single search results. As opposed to mere preview displays [Ru08], parts of the text in that area can be selected and used as a basis for additional query restrictions. The algorithm determines which aspects are most likely addressed in the selected text range, and what statements are made about them in the selected search result. It then suggests to add additional filters to obtain only documents that express the same sentiment about the mentioned aspects.

An area at the bottom of the visualization displays a pie chart indicating how many documents of the total set match the filter criteria. If the documents of the total set can be clearly divided into distinct subtopics, there will be several pie charts that can be compared to each other. Figure 1 is shown with the exemplary use case of filtering restaurant reviews, so the bottom area shows one pie chart per restaurant (in an actual implementation, restaurant names would be shown instead of placeholders). The radius gives a hint about the total relative number of reviews per restaurant, and the pie charts are first sorted by the percentage of reviews matching the filter criteria, then by the number of reviews.

4 Example of Application

In the following, we present a usage scenario that demonstrates the usefulness and applicability of Aspect Grid. Alice wants to find a suitable restaurant for seeing a friend. By entering the first few search terms such as *restaurant*, *recommendation* and the name of the city, the search engine recognizes that the user is interested only in restaurant reviews. Therefore, the Aspect Grid visualization is loaded with aspects related to restaurant reviews. The user starts to check some of the lines in the table and notices that some items need to be filtered out. She notes that it is important for her that the food has left a positive impression on the reviewers, or that reviewers had a neutral impression of the food, in

which case they should have judged the price as positive ("Either the food must be extraordinary, or if it isn't, it should at least not cost much."). She adds these three filters—the state right before adding the filter on the *Price* aspect is depicted in Figure 1 (left)—and connects them with the appropriate *any* and *all* operators. After selecting and reading some of the filtered reviews, she discovers a statement about the presentation of the food in one of the reviews and realizes that she would appreciate that aspect, as well. Therefore, she highlights the statement in the review text, whereupon a positive judgment of the *Presentation* aspect is automatically suggested as a new filter as depicted in Figure 1 (right). She adds the filter by clicking on the suggestion. The bottom area of the Aspect Grid display now shows a short list of restaurants, ordered by the ratio of reviews that matched her filters compared to the reviews that did not match her filters. Thus, based upon her custom aspect-related review filters, she picks one of the first few restaurants.

While this usage scenario shows a specific application example for Aspect Grid, we are confident that our approach is useful for iteratively refining filters on any collection of text documents from which different aspects can be extracted.

5 Conclusion and Future Work

We have presented an interactive visualization concept that allows for the filtering of documents based upon aspects that were extracted from unstructured text. Different from the related works, our approach incorporates the unstructured documents from preliminary result sets into the visualization in a way that they can be used as a basis for suggestions on additional filter restrictions. This way the user can build the desired filter more intuitively and efficiently and thereby achieves his or her intended goal faster.

We are planning to implement a prototype of the visualization and evaluate the usefulness of the approach in a user study. Possible extensions of the concept that may be recognized as necessary in user experiments include an enrichment of the data extracted from unstructured text with structured data, as well as an implicit training method for enhancing the aspect recognition precision in documents on the fly.

References

[AKG10] Aula, A.; Khan, R. M.; Guan, Z. How Does Search Behavior Change as Search Becomes More Difficult? In *Proc. CHI '10*, pages 35–44, New York, NY, USA, 2010. ACM.

[BE10] Brody, S.; Elhadad, N. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *NAACL HLT '10*, pages 804–812, Stroudsburg, PA, USA, 2010. ACL.

[Bo10] Bozzon, A. et al. Liquid Query: Multi-domain Exploratory Search on the Web. In *Proc. WWW '10*, pages 161–170, New York, NY, USA, 2010. ACM.

[Bo11] Bozzon, A. et al. Visualization of Multi-Domain Ranked Data. In *Search Computing*, volume 6585 of *LNCS*, pages 53–69. Springer, Berlin/Heidelberg, Germany, 2011.

[Do08] Dork, M. et al. VisGets: Coordinated Visualizations for Web-Based Information Exploration and Discovery. *IEEE TVCG*, 14(6):1205–1212, 2008.

[Ha06] Hansaki, T. et al. FindFlow: Visual Interface for Information Search Based on Intermediate Results. In *Proc. APVis '06*, pages 147–152, Darlinghurst, Australia, 2006. ACS.

[He09] Heim, P. et al. RelFinder: Revealing Relationships in RDF Knowledge Bases. In *Semantic Multimedia*, volume 5887 of *LNCS*, pages 182–187. Springer, Berlin/Heidelberg, Germany, 2009.

[Hi11] Hindle, A. et al. Clustering Web Video Search Results Based on Integration of Multiple Features. *World Wide Web*, 14(1):53–73, 2011.

[Iw12] Iwata, M. et al. AspecTiles: Tile-Based Visualization of Diversified Web Search Results. In *Proc. SIGIR '12*, pages 85–94, New York, NY, USA, 2012. ACM.

[Jo98] Jones, S. Graphical Query Specification and Dynamic Result Previews for a Digital Library. In *Proc. UIST '98*, pages 143–151, New York, NY, USA, 1998. ACM.

[Ko09] Koch, S. et al. Iterative Integration of Visual Insights During Patent Search and Analysis. In *VAST 2009*, pages 203–210, 2009.

[Li12] Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Ma11] Marcus, A. et al. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proc. CHI '11*, pages 227–236, New York, NY, USA, 2011. ACM.

[MP13] Mirylenka, D.; Passerini, A. ScienScan—An Efficient Visualization and Browsing Tool for Academic Search. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *LNCS*, pages 667–671. Springer, Berlin/Heidelberg, Germany, 2013.

[MRS08] Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, Cambridge, UK, 2008.

[PL08] Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Ra05] Radev, D. et al. NewsInEssence: Summarizing Online News Topics. *Commun. ACM*, 48(10):95–98, 2005.

[RTB11] Rástočný, K.; Tvarožek, M.; Bieliková, M. Supporting Search Result Browsing and Exploration via Cluster-Based Views and Zoom-Based Navigation. In *Proc. WI-IAT '11*, volume 3, pages 297–300, 2011.

[Ru08] Russell, A. et al. NITELIGHT: A Graphical Tool for Semantic Query Construction. In *Proc. SWUI '08*, volume 543. CEUR-WS, 2008.

[Sh94] Shneiderman, B. Dynamic Queries for Visual Information Seeking. *IEEE Software*, 11(6):70–77, 1994.

[Sp93] Spoerri, A. InfoCrystal: A Visual Tool for Information Retrieval & Management. In *Proc. CIKM '93*, pages 11–20, New York, NY, USA, 1993. ACM.

[St12] Stab, C. et al. Semantics Visualization for Fostering Search Result Comprehension. In *The Semantic Web: Research and Applications*, volume 7295 of *LNCS*, pages 633–646. Springer, Berlin/Heidelberg, Germany, 2012.

[Ye03] Yee, K.-P. et al. Faceted Metadata for Image Search and Browsing. In *Proc. CHI '03*, pages 401–408, New York, NY, USA, 2003. ACM.