Visualization for Relevance Assessments

Omar Alonso
Microsoft
omalonso@microsoft.com

Abstract

Visual data exploration is a well-known technique in the data mining toolbox for analyzing and understanding large amounts of data. In information retrieval, visualization is still perceived as a front-end or user facing feature while, at the same time, it can be very effective in the back-end for data analysis. Currently, the research community does not sufficiently analyze the reasons underlying the assessments they use for evaluation, which may lead to bias or error when reporting system performance. In this paper we examine the integration of information visualization techniques for visually exploring and analyzing relevance assessments. We argue that visual tools should be integrated as a core component of the infrastructure for relevance assessment and evaluation.

1 Introduction

In this new era of increasingly large data sets, researchers and practitioners in the areas of information retrieval (IR), machine learning and natural language processing (to name but a few) rely heavily on collecting high quality labels during research and development. Feedback from labels is a fundamental element of training set generation, evaluation and experimentation efforts. Labels, also known as assessments or preferences, are provided at some point in the process, by humans.

Relevance assessment is at the core of most modern information retrieval approaches. Since IR data sets have grown in size, it is impractical for humans to label an entire collection. Instead, representative parts of the collection are sampled for human labeling, and predictive features are identified based on this subset so that more scalable machine learning approaches can be used to label the remaining part. It is essential to focus on eliciting reliable high quality labels during the early stage of the process.

The process of gathering relevance assessments in an effective manner is difficult and time intensive. Even with the rapid adoption of crowdsourcing techniques, many issues remain open. To gather useful results, many factors must be considered. Clear instructions, user interface guidelines, content quality, inter-rater agreement metrics, work quality and worker feedback are important characteristics of a successful experiment, regardless of whether the judges providing the labels are experts or

crowdsourced workers. Furthermore, designing and implementing experiments that require thousands or millions of labels is very different from conducting small scale research investigations.

Once all the labels have been collected, the focus is on the performance of a specific metric or any other measurable outcome that helps quantify and compare results. In practical terms, this means storing the labels in a database so different reports can be produced at any moment. While achieving a given metric target is usually the main outcome, the exploration of the labels in context, in this case documents, is usually a secondary task.

Several factors influence relevance judgment. When the experiment is not positive, typical first reactions are either to blame judges for not doing their job properly, or, question whether the task instructions are clear. As we will see later, relevance assessment is hard and multiple considerations may influence the outcome of an experiment. Exploring and analyzing assessment data can be useful for learning more about the relevance judgment process.

Information visualization for IR is an area of research with an emphasis on alternative user interfaces for exploring and assisting search scenarios. Unfortunately, the use of visualization in the search domain has been limited, most likely due to the popularity of the prevailing so-called "10-blue links". Visual analytics incorporates information visualization techniques with a focus on data analysis tasks. The idea is to mine, explore and inspect data sets with the aid of visualization tools. The advantage of visual data exploration is that the researcher is directly involved in the mining process and not a mere consumer of the presentation [11]. This cross-disciplinary interaction can benefit both IR and visualization communities [1].

We think that it is possible to use similar visualization techniques to explore and gather better insights into how and why judges assign labels, and simultaneously gauge if the experiment is sound. Our goal is not to propose yet another visualization metaphor – but rather, to use the many visualizations techniques already available as an integral part of the relevance assessment collection process and evaluation lifecycle. In the rest of this paper, we describe how this integration can be done by showcasing a working example.

2 Challenges with Relevance Assessments

Often perceived as very laborious and a hassle, the process of collecting relevance assessments is difficult and time consuming. At a very high level, the practice involves sampling a set of documents and creating an experiment which instructs judges how to evaluate the relevance of those documents on a given scale according to a topic or query. Judges inspect each document and assign a label.

We assume that judges are honest and will try to provide the best answer possible. However, it is always desirable to have quality control mechanisms to detect spammers and poor work. Even experts can make mistakes. Due to the visual nature of the task, it is possible that something caught the attention of a judge and might have influenced the answer. Capturing an exact copy of what the judge saw is a good practice for investigating the results in more detail.

All human tasks have a level of difficulty, and relevance assessment is no exception. Work by Villa and Halvey on difficulty shows that relevant documents require more effort to judge compared to highly relevant and non-relevant documents, and that effort increases as document size increases [15]. The more subjective the task is, the more difficult it is for workers. Sometimes increased effort

is due to the complexity of the experiment, or unfamiliar content for the assigned judges. Different assessors produce different relevance judgments. Al-Harbi and Smucker propose four categories for classifying differences between primary and secondary assessors [2]. Further related work on assessments in general includes [13] and [14].

Traditionally, document relevance assessments experiments have not changed much over the years. That said, for new sources like social data or for more complex evaluation experiments, there may be more than just one step for collecting labels. For example, a combination of different crowds and stages for collecting labels in near-duplicate news evaluation is presented in [4].

Debugging crowdsourcing assessments tasks can also be a challenge. Work, workers, and task design are contingent on one another and adjusting one element may have a substantial effect on the others. The framework presented in [3] explores the nature of the task and ways to ensure that reliable labels have been assigned to a scalable subset of the larger dataset.

After the experiment has finished, we are interested in exploring the results visually and looking at the data from many different angles.

3 Data Visualization for Relevance Assessments

The field of information visualization focuses on the use of computer-supported, interactive, visual representations of abstract data to amplify cognition [5]. This is an area of active research with contributions from academia and industry. There are existing visualization products, add-on features and open source toolkits that allow the development of visualization applications very quickly.

As mentioned earlier, information visualization, thus far, has not become widespread in the search domain. Hearst's book on search user interfaces describes the challenges of visualizing textual data and presents a number of application examples in different parts of the search process [9]. Hearst suggests that visualization is a promising tool for the analysis and understanding of text collections.

There has been little work on methods to visually explore relevance assessments. Two recent examples are Hauff's visualizations on system effectiveness based on TREC data using D3[1] [7, 8] and the visual representation of search sessions proposed by Cerviño Beresi *et al*. [6].

The idea is not to use a visualization metaphor to present the assessments results because it looks pleasant to the eye. Rather, use visual data exploration as part of the overall process for gathering assessments and improving label quality. Similar to the objectives in visual analytics, the aim is to both detect expected patterns and discover unexpected ones [11]. We believe that there is a lot of potential to be gained when exploring assessment data sets with visualization techniques.

4 Relevance Assessments Visual Exploration

In this section we describe how such integration would work using a simple example that is easy to replicate. We assume access to a system for gathering labels, such as a crowdsourcing platform like CrowdFlower or Amazon Mechanical Turk. The procedure for our working example is as follows. We run a crowdsourcing experiment and download the results after completion. Taking the raw set of

[1] http://D3js.org

labels provided by the workers as input, we transform the data so it can be read by a visualization tool for data exploration.

4.1 A Simple Experiment

Our relevance experiment consists of comparing the results of two search engines, named A and B. For the purpose of this experiment, we simulate two different ranking functions using the Bing API[2], which given a query will return the top 10 web results. Using the 10 items from the list, we create two sub-lists: A (positions 1 to 6) and B (positions 7 to 10). We de-brand the results and present a simple task that requires the worker to answer which search engine is better: "A", "B", or "Same". We also randomize the placement of A and B so it is not obvious to identify control from treatment. As a basic quality control mechanism, we include a few honey pots for checking worker performance. Figure 1 shows the human intelligence task in Mechanical Turk.

Figure 1. A comparison experiment of two search results.

4.2 Facets

Usually a ranking function contains many different parameters that are tuned and monitored by running online and offline experiments. Queries also enclose many attributes. We can use some of this metadata as facets for organizing the assessments, allowing us to explore the data set by applying different filters.

We do not claim that faceted browsing is the only technique for visually exploring a data set. There are many different ways of organizing data; faceting is one common approach. We demonstrate the integration of relevant assessment data with facets using two systems: Pivot and Exhibit.

4.3 Pivot

Pivot makes it easier to interact with massive amounts of data by visualizing thousands of related items at once. The visualization metaphor uses images as the main object that allows the display of high-resolution content without long load times, while the animations and natural transitions provide context and prevent users from feeling overwhelmed by large quantities of information. Pivot is available as part of Silverlight[3] or as a standalone application [12].

[2] http://msdn.microsoft.com/en-us/library/dd877956.aspx
[3] http://www.microsoft.com/silverlight/pivotviewer/

Figure 2 shows how the data from the experiment looks in Pivot. The visualization contains three panels: the filter panel, the main content in the middle, and, to the right, the details per item (in our case document assessments). As facets to sort, filter or group the collection we have the worker identification, work time, query, query length, query type, entity type on query, event, data set selection, A, B, and the label provided by the worker. These facets were defined for this specific experiment and it is expected they are domain dependent.

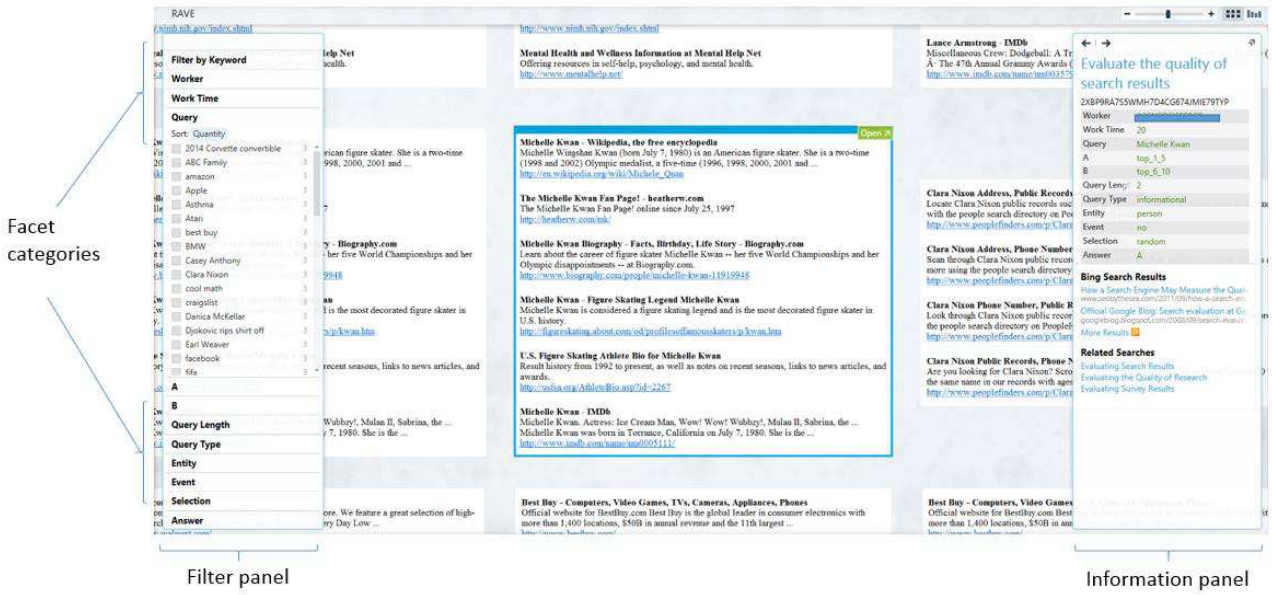

Figure 2. Exploring relevance assessments with Pivot.

When we zoom in on a collection item, detailed information about it appears in the information panel. This part of Pivot serves two purposes: providing detail on the assessment and encouraging exploration, for example links to find similar or related items. It is also possible to change the view, in this case to a graph in order to visualize the distribution of assessments given the many filters. Figure 3 shows an example that favors search engine A.

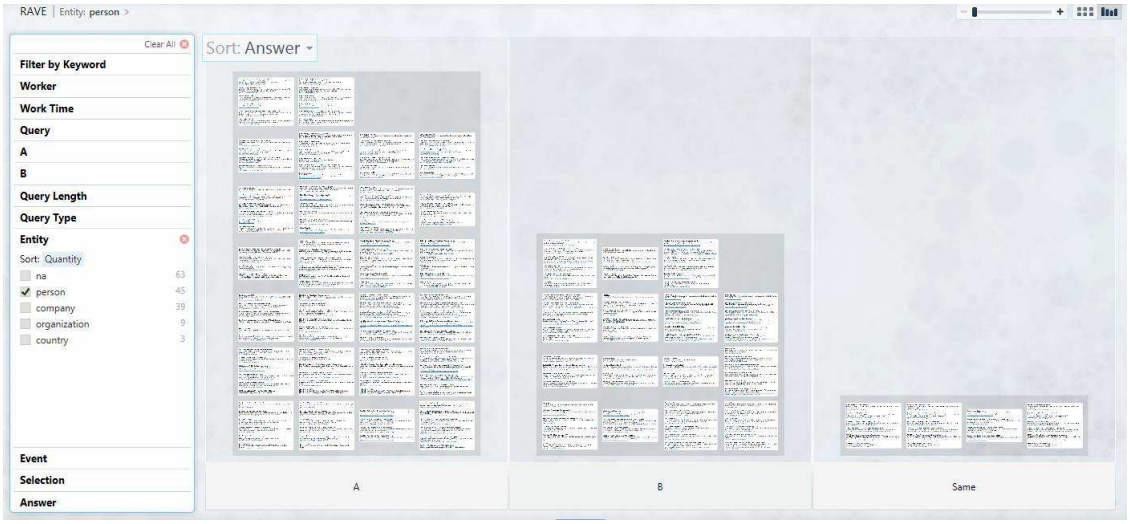

Figure 3. Analyzing preferences for queries that contain the name of a person.

Because Pivot relies on images, it is necessary to create a collection of images first. Each assessment, in our case the list of search results, was captured as an image and associated with all the metadata. A simple script transforms the raw data from MTurk into the cxml format, making it available for Pivot.

4.4 Exhibit

Exhibit[4] is a lightweight publishing framework that allows developers to create web pages with advanced text search and filtering features, including visualizations. A nice feature of Exhibit is that it does not require a lot of development knowledge to produce interactive pages that exploit the structure of a data source for better browsing and visualization [10].

Using the same data set as in the previous example and with a similar approach, we show how to explore the assessment data set with Exhibit. Figure 4 shows how the data from the experiment looks with Exhibit. In contrast to Pivot that has a predefined layout, we can design our exhibit in a more flexible way. In our case, we center the document thumbnails on the page and provide the facets, workers and assessment answer on the left, and query type, query length, and entity type on query on the right. The example presents the assessments sorted by entity type.

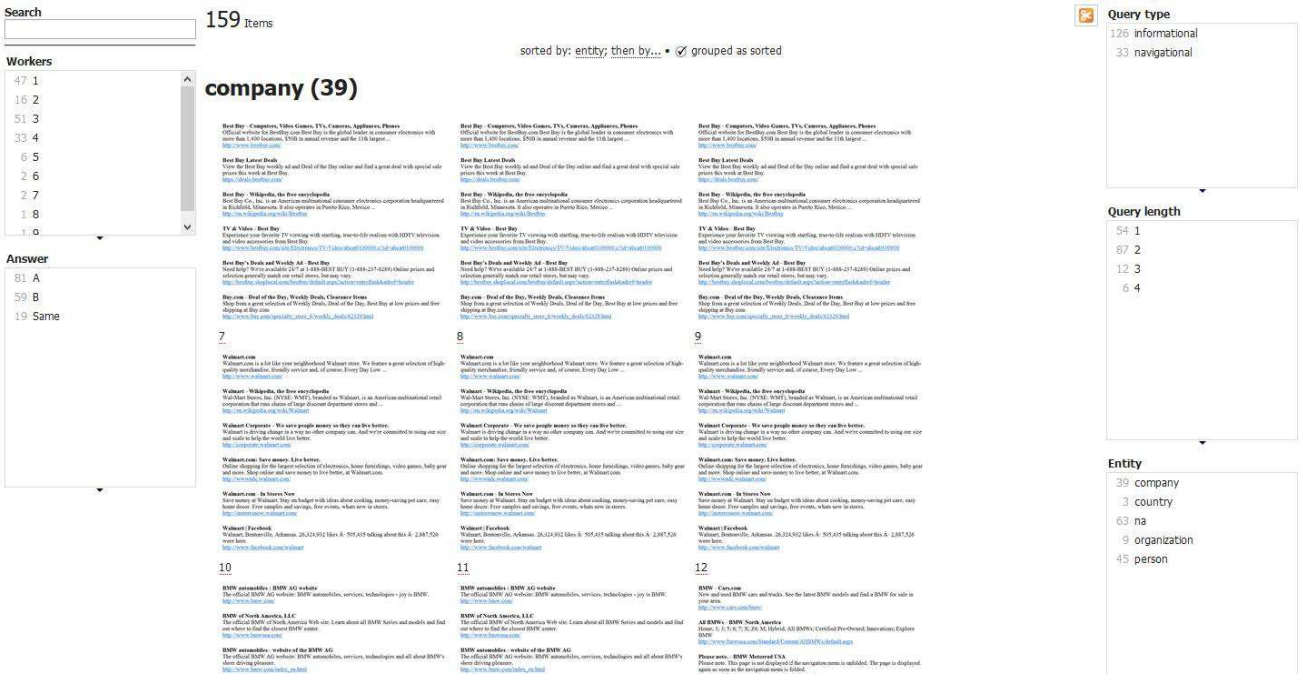

Figure 4. Exploring relevance assessments with Exhibit.

If we click on the link next to the thumbnail, we can explore the document in more detail (Figure 5).

[4] http://simile-widgets.org/exhibit/

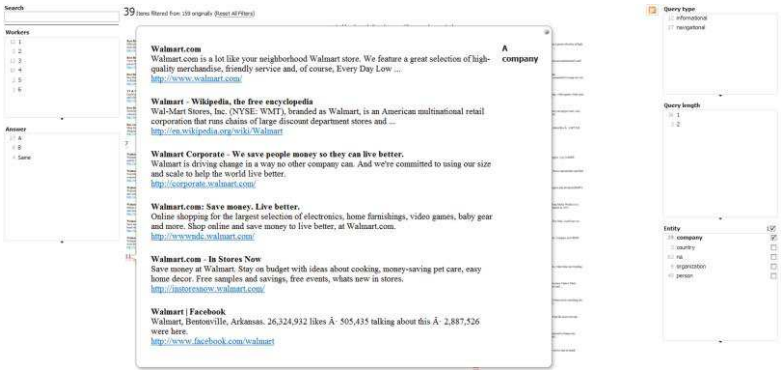

Figure 5. Exploring the search results document with Exhibit.

Constructing an exhibit involves two main tasks: providing the data and authoring the presentation. In our example, a simple script processes raw data from MTurk to generate the necessary data in JSON format. The data model of each exhibit is a set of items in which each item has a type and several properties. The web page authoring involves creating some basic HTML code.

Using Exhibit has two benefits: there is no installation required and no server side programming – everything is done in the browser, and agile development for more interactive exhibits.

5 Conclusions

In this paper we investigated how visualization techniques can be used as part of the relevance assessment process to help gather better insights into label quality and experimentation in general. We demonstrated how a couple of faceted-based visualization techniques can assist in the exploration of relevance assessments. We believe that data visualization should be part of the standard instrumentation and infrastructure for relevance assessments and evaluation. In terms of engineering effort, the integration work required is modest.

Pivot and Exhibit are examples of visual data exploration tools, and there are many more metaphors that can also be tested. At the same time, there are opportunities for designing new visualization metaphors for relevance assessments. An open issue is to evaluate how effective visualization techniques can be in the relevance assessment and evaluation.

While it is true that the user audience for visual analytics is small, better insights from labels can lead to infrastructure improvements that can translate into a better search experience for a much larger user base.

All the source code for the examples, including the raw data and images, is available from GitHub (https://github.com/oalonso/rave).

6 Acknowledgements

We thank Stewart Whiting and the Forum editors for helpful comments.

7 References

[1] Maristella Agosti, Nicola Ferro, Pamela Forner, Henning Müller, and Giuseppe Santucci (Eds.). *Information Retrieval Meets Information Visualization*. Springer, 2013.

[2] Aiman Al-Harbi and Mark Smucker. "A Qualitative Exploration of Secondary Assessor Relevance Judging Behavior". In *Proc. of IIIX* 2014, 195-204.

[3] Omar Alonso, Catherine Marshall, and Marc Najork. "Crowdsourcing a Subjective Labeling Task: A Human-Centered Framework to Ensure Reliable Results", MSR-TR-2014-91, http://research.microsoft.com/apps/pubs/default.aspx?id=219755

[4] Omar Alonso, Dennis Fetterly, and Mark Manasse. "Duplicate News Story Detection Revisited". In *Proc. of AIRS* 2013, 203-214.

[5] Stuart Card, Jock Mackinlay, and Ben Shneiderman. *Readings in Information Visualization. Using Vision to Think*. Morgan Kaufmann, 1999.

[6] Ulises Cerviño Beresi, Yunhyong Kim, Dawei Song, and Ian Ruthven. "Why Did You Pick That? Visualising Relevance Criteria in Exploratory Search". *Int. J. on Digital Libraries* 11(2), 2010, 59-74

[7] Claudia Hauff. "Ranking Retrieval Systems without Relevance Judgments". http://www.st.ewi.tudelft.nl/~hauff/visualization/trecVis.html

[8] Claudia Hauff. "The Uniqueness of QREL Contributions". http://www.st.ewi.tudelft.nl/~hauff/trec_vis/trecQrelVis.html

[9] Marti Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[10] David Huynh, David Karger, and Robert Miller. "Exhibit: Lightweight Structured Data Publishing". In *Proc. of WWW* 2007, 737-746.

[11] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. "Visual Analytics: Definition, Process and Challenges". In A. Kerren *et al*. (Eds.) *Information Visualization*. Springer, 2008.

[12] Microsoft Pivot. http://research.microsoft.com/en-us/downloads/dd4a479f-92d6-496f-867d-666c87fbaada/default.aspx

[13] Ian Ruthven, Mark Baillie, and David Elsweiler. "The Relative Effects of Knowledge, Interest and Confidence in Assessing Relevance". *Journal of Documentation*, Vol. 63, No. 4, 2007, 482-504.

[14] Jean Tague. "The Pragmatics of Information Retrieval Experimentation" in K. Spärck Jones (Ed.), *Information Retrieval Experiment.* Butterworths, 1981.

[15] Robert Villa and Martin Halvey. "Is Relevance Hard Work? Evaluating the Effort of Making Relevant Assessments". In *Proc. of SIGIR* 2013, 765-768.