Information Services & Use 36 (2016) 57–63
DOI 10.3233/ISU-160793
IOS Press

Visual information discovery

Thomas Grandell
CEO, Etsimo Ltd, Puutarhakatu 31 C 12, 20100 Turku, Finland
E-mail: thomas.grandell@etsimo.com

Abstract. Information is only valuable when it can be found. It is relatively easy to find information that we know exists; you just need to know what you want to find, where to look for it, and how to express yourself. It might require a little effort to find it, but this work is usually rewarded. But, what if you're looking to acquire new knowledge or want to search for unusual findings? Normally you start with what you know and learn as you go and in the end arrive at some conclusion, but how do you know you have found all relevant information and how do you know how this information is connected to the rest of the information space? The answer up until now is that you don't, and to get the big picture you would have to aggregate your results manually. Microsoft has studied search behavior and a big part of the daily searches are exploratory in nature, where the users aren't sure what they want to find and are struggling while trying. The Etsimo Visual Discovery engine offers an alternative, or complement, to lookup searching by providing a transparent and user-driven way to visually navigate the information space, learn as you go, and find relevant information.

Keywords: Information discovery, exploratory search, visualization

1. Introduction

We at Etsimo believe that relevant information should be available to everyone, at all times and everywhere. However, today's world presents us with a plethora of challenges to make this possible; data deluge and information overflow, silos, information bubbles, and lack of tools for effectively navigating the information space. Our mission is to battle these challenges. We help people learn new things by making it easy for them to find, or discover, relevant content in any data.

To make it possible for people to perform exploratory searches, we've developed a cloud based (SaaS) visual discovery platform. Our patent-pending technology, formerly known as "SciNet," is a result of three years of research at the Helsinki Institute of Information Technology, HIIT, which is a joint institute of the University of Helsinki and Aalto University, Finland, and there are several scientific papers published about it.

With this solution, we combine the efforts of man and machine to an iterative discovery process with the help of a knowledge support system. The core of this is combining three things; 1) artificial intelligence and machine learning to build 'intent models' predicting the user's current and future search intents, 2) visualizing the results, i.e. intents on the user's screen, and 3) making it possible for the user to give feedback to the system to refine the search, which leads us back to recalculating the intent models in step 1. So, the machine, which is really, really fast at sifting through millions of documents in milliseconds and at performing staggering amounts of calculations in less than a blink of an eye, does exactly that and presents the results to the human. The human, who is superior in logical reasoning and in identifying interesting topics in the information space does exactly that, and gives feedback to the machine. This is a two-way reinforcement learning process where the machine is teaching the

human about the information space, and the human is teaching the machine about his or her search intent. Man and machine are working seamlessly together towards a common goal to discover relevant information.

2. Traditional and exploratory searching

Before we dive deeper into how this works, it is important to understand the differences between 'traditional' or lookup searching and exploratory searching. Lookup searching is what we do when we use the query-response model to find information. These models are in use everywhere, and the most known implementations are Google, Bing, Yahoo, Yandex, and Baidu. The user enters a few keywords, 'the query,' and presses a button to get the results, 'the response.' If the user isn't satisfied with the results, he or she has to improve the original query to get better results. So, the user's search intent is captured in the initial query only, and there are no possibilities of refining the results within the query. The 'normal' way is to learn as you go, by examining the results to learn more and to identify elements and keywords that might lead us closer to the desired information. Another characteristic of the query-response model is that we receive long result lists spanning several pages and millions of hits, with the most 'important' ones at the top of the list. This 'importance' is determined by several different factors, i.e. popularity (what everybody else is searching for/clicking on), previous searches, geographical location including culture, and search engine optimization (SEO). This works really well when we're using the search engine as an extension of our human brain, i.e., we remember the name of our favorite pizza place and we use that to search for their website or phone number to make a reservation. The ranking of the results based on popularity have become so good that we seldom have to go beyond the first couple of pages to find what we are looking for and almost never have to use the complicated advanced search features. We could say that these work extremely well when you know 1) what you are looking for, 2) where to look for it, and 3) how to formulate a query to find it. If these three things are within your knowledge, you almost always find what you are looking for by doing a little work.

Lookup searching works less well when you want to learn new things and acquire new knowledge, or when you're searching for unusual findings, unknown unknowns. I believe Donald Rumsfeld is the one that made this expression known. He stated that we normally know what we know and even know what we don't know, but sometimes we stumble upon information that we didn't even know we didn't know, and these are the unknown unknowns.[1]

So why doesn't lookup searching work well in these cases? First, there is an unfamiliarity with the domain or topic in question – there is a gap between our current knowledge and the desired knowledge, and we don't know exactly what we're looking for. This leads to the second challenge; it's hard to formulate a good query (= one that produces the relevant hits) when you don't know the domain or topic. Normally, the results consist of either thousands of hits (= impossible to go through them all) or zero hits. This is commonly known as a vocabulary mismatch problem, which means that the vocabulary of the person trying to locate the information is different from the vocabulary of the person (or machine) that put the information into the system. Some information providers try to eliminate the second problem, namely the thousands of hits, by applying filters or facets to the results, enabling the user to refine his or her search by drilling down into the information in a controlled manner. This leads to the third challenge; the user gets trapped within the initial query. For each iteration (e.g. applied filter), the user gets a smaller

[1] D. Rumsfeld and R. Myers, DoD news briefing, February 12, 2002: http://archive.defense.gov/transcripts/transcript.aspx?transcriptid=2636 (last checked, June 19, 2016).

and smaller subset of the initial query, thus limiting the possibility to discover unexpected results and connections in the content. The fourth challenge is about the big picture. For each search, i.e. new query, the user gets new results and aggregating all the relevant results from a search session consisting of several queries requires manual aggregation. This makes it hard to get a holistic view of the topic or domain, and any such manual aggregation gets old really fast. On top of this, there is the uncertainty about if all the relevant information has truly been found. The last, or fifth challenge is related to the way information is ranked in today's search solutions. They are 'black boxes,' where the user has no way of knowing why the displayed results are displayed. As stated earlier, we know that the results are based on popularity and such, but the exact logic behind the algorithms producing hits is proprietary information. These algorithms create so called information bubbles, where the search provider is in control of what information is shown. This is a real challenge when you want to find novel information – it can, of course, be found, but you need to know it exists, and also more or less the exact keywords to use to find it. Microsoft did a study on the search logs of their Bing search engine back in 2014, and they concluded that up to half of all five billion daily Internet searches are exploratory; i.e., the users are searching out of curiosity with the aim of learning new things. Currently, the mainstream search solutions don't support this very well, as, according to the same study, 59% of the users exploring did so struggling.

3. Exploratory searching

To really prove my point consider the following. Suppose you're an expert in domain X and want to learn something new. For anything you know, you can build a query. But how do you query for something you don't know about? Using your normal keywords and queries to go through several thousands of hits to see if there is something new to learn, is hardly a solution.

The Etsimo visual discovery solution doesn't compete with lookup searching, as the use case for exploratory searching is different. Our solution uses artificial intelligence and machine learning to build session-specific intent models on-the-fly that are tailored to the user and session-based on the user's behavior. In Fig. 1(left) below, the user has initiated a search by using the keyword 'jobs' with the intent to find information about Steve Jobs. The system builds the current search intent based on this initial query in the light striped area of the intent radar. The keywords related to the current search intent are black on the radar. In addition to the current intent, we also predict several future intents, i.e. future

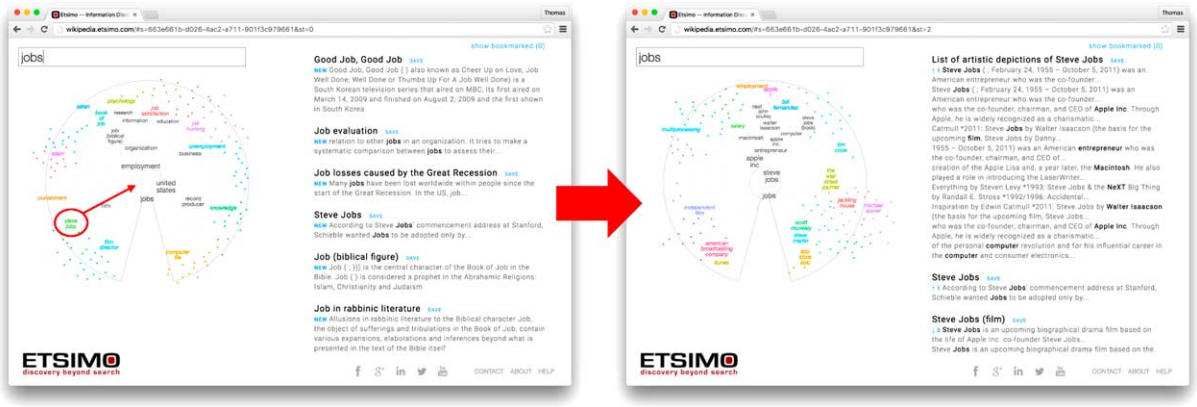

Fig. 1. Using the radar to refine search results.

search directions. These are clustered around the current intent, and each topic area or intent has its own color, with the strongest keyword readable while the rest of the keywords in the cluster minimized to dots. If the user places the mouse pointer over a cluster, the additional keywords are zoomed in and individually selectable.

Note from Fig. 1(left) that as 'jobs' is an ambiguous word, the search results reflect this ambiguity as the machine is giving a cross-section of the information space. As the user examines the keywords on the radar, he or she notices 'Steve Jobs' around 8 o'clock on the radar. The user tells the machine to move the results towards 'Steve Jobs' by dragging the keyword pair to the center of the radar. This action immediately triggers a recalculation of the intent models and the results, and this is shown on the right side of Fig. 1. Now the intent models are built mainly around 'Steve Jobs' instead of just 'jobs' and the results are mirroring this. Keywords can also be dragged out of the radar for the inverted effect, i.e. telling the machine that we're definitely not interested in anything containing that keyword.

When refining the intent models, the machine remembers the initial search query for about eight to twelve moves, depending on how 'curved' the user's road through the information space is. This means that when we're starting with 'Tesla,' and go from that to 'Elon Musk,' the machine ranks [Elon Musk + Tesla] higher than just [Elon Musk], with the same also being true for the combinations [Elon Musk + SpaceX] and [Elon Musk + SolarCity]. This makes the whole search and discovery process completely transparent. The machine is communicating with the human through the radar by showing how the keywords are influencing the result, and the human directs the search by giving feedback.

The demo environment used in the example above utilizes the complete content of the English version of Wikipedia, some five million articles resulting in around thirty-five million documents in our index. It can be found at wikipedia.etsimo.com. The interface used in the demo is built to demonstrate the capabilities of Etsimo's Visual Discovery Engine (VDE). The UI operates against a frontend API, which means we can use any kind of interface that supports giving feedback on the presented search results back to the machine for recalculation.

The discovery engine itself is completely data-driven. This means that we don't need any special structure in the data to be able to utilize the VDE on top of it. So, no taxonomies, no ontologies, no hierarchies, and no knowledge graphs – everything is structured on-the-fly using artificial intelligence and machine learning to build the intent models. To operate, we need only keywords in a full-text index, and for the Wikipedia demo we use four different data types; 1) caption, 2) body-text, 3) keywords, and 4) URL to the original Wikipedia article. An example data structure for a document in .json format could look like this in its simplest form:

```
{
 "url": "http://example.com/item1",
 "kwds": ["Keyword one", "Keyword two"],
 "kwd_frq": [1, 2],
 "title": "The title",
 "text": "The long body text",
 "term_count": 3,
 "date": "yyyyMMdd",
 "language": "eng",
 "source": "the source"
}
```

```
kwds contains an array of keywords while
kwd_frq contains the keyword count in the document in
question. Given document

title: 'Test Document'
text: 'My sample test document content.'

the keywords and frequencies could be
kwds: ['Test', 'Test Document', 'Content']
kwd_frq: [2, 1, 1]
```

Of course the data can be enriched with additional metadata, e.g. 'Publishing Date' and 'Author,' to enable greater manipulation and structuring of the results. The lack of requirements for data structure makes the Etsimo solution perfect for combining different types of content, i.e., breaking silos, without having to start with a data cleaning and structuring project. If the data can be delivered in a format adhering to the example above, using our standard components and configurations, we can have a test environment operating on your data up and running in a matter of hours. The behavior of the VDE is completely configurable, so we can, for example, balance the amount of exploitation versus exploration. Exploitation is when the machine is most focused on finding information closely related to the initial search. Exploration again means that the machine also presents the user with results a little bit further away from the initial search, enabling more possibilities for discovery.

4. Usability and performance tests

We wanted to make sure visual discovery delivers in real life, too, and not just in our heads. To verify this, we've arranged several usability and performance tests with real, unbiased users. For the test described below, we wanted to verify two things, 1) retrieval performance; i.e., to measure the effectiveness of the system, namely how accurate are the results that the system is able to return in response to user interactions, and 2) user performance; i.e., to measure the quality of answers the users provided as responses to a given task when using a given system. The details of this study can be found in the scientific publication listed in the additional reading section at the end of this document.

To measure the different performances, we benchmarked visual discovery against text-based searching on the same content. For this, two post-doctoral researchers from the domains of information retrieval and machine learning were recruited as experts to define the task. The experts wrote task descriptions using the template 'Imagine you're writing a scientific essay on the topic of semantic search/robotics. Search for scientific documents that you'd find useful for this essay.' Twenty researchers were recruited as test subjects, and split into two groups according to their area of expertise. These two groups were then split into two sub-groups; one that would use the visual discovery interface and one that would use a text based search interface, and then trained on the system they were to use in the tests. The data set used contained fifty-plus million scientific publications, including the whole content of Thompson Reuters' Web of Science and the digital libraries of Springer, ACM and IEEE.

The test subjects were given thirty minutes to complete the tasks and we logged all interactions with the systems and all the articles and keywords presented by the system in response to these interactions. Three measures – precision, recall, and F measure – were defined and we also conducted post-task interviews with the participants. The results were encouraging. Figure 2 below summarizes the results,

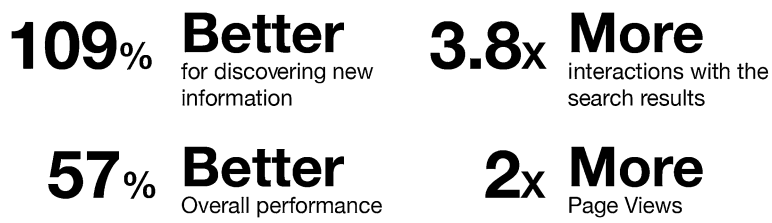

Fig. 2. Novel and overall performance in the usability test.

and from the research paper listed at the end of this article, we can see that for *novel* information retrieval performance, the visual system pulls ahead at a very early stage and the gap between the two widens the longer the test progresses. The curves for *overall* performance shows that the systems perform equally well for around six minutes before the visual system starts to pull away. This tells us that we humans manage to exhaust our stockpile of obvious keywords fairly fast, and therefore need to start reading the hits to find more useful keywords, which slows down performance and leads to less comprehensive results when using a text-based search interface. Identifying suitable keywords from a given visualized set is cognitively much easier and we can focus on the information instead of focusing on finding new keywords to operate the system with.

5. Conclusion

In conclusion, information only has value when it can be found. Lookup searching is very effective when you know what you're looking for, and as exploratory searching and lookup searching have different use cases, they don't compete with each other, but rather complement each other. So to make sure information is found, one should offer appropriate tools on top of one's content, as using appropriate tools for exploratory searching gives superior results.

Better findability can also be turned into more revenue; in our tests, we found that presenting relevant content to the user from the start immediately engages the user on a much higher level and he/she starts to explore and interact with this content. So, more relevant content leads to longer stays and more pages viewed. Relevant content also leads to a higher trust factor with regard to finding everything there is to find, and this, in turn, is a key factor for higher conversion rates and returning users. After all, it is only human to choose the least complicated way to achieve the desired result – if you know something delivers, why wouldn't you use it?

About the author

Thomas Grandell is the CEO and co-founder of Etsimo Ltd, where he is currently working on a new and unique visual discovery technology that revolutionizes the way we consume and interact with information.

References

[1] Etsimo's Wikipedia demo at wikipedia.etsimo.com. See demo at: http://wikipedia.etsimo.com/.
[2] D. Glowacka, T. Ruotsalo et al., Directing exploratory search: Reinforcement learning from user interactions with keywords, in: *ACM IUI 2013 Proceedings of the 2013 ACM International Conference on Intelligent User Interfaces*, Santa Monica, CA, USA, March 19–22, 2013.

[3] A. Hassan, R.W. White et al., Struggling or exploring? Disambiguating long search sessions, in: *ACM WSDM 2014 Proceedings of the 2014 ACM Conference on Web Search and Data Mining*, New York City, NY, USA, February 24–28, 2014.
[4] T. Ruotsalo, G. Jacucci et al., Interactive intent modeling: Information discovery beyond search, *Communications of the ACM* **58**(1) (2015), 86–92.