# Interactive Topic Analysis with Visual Analytics and Recommender Systems

**Eduardo Veas**[1,2]
Institute for Information Technologies and Communication[1]
National University of Cuyo

**Cecilia di Sciascio**[2]
Knowledge Visualization[2]
Know Center GmbH

**Abstract**

The ability to analyze and organize large collections, to draw relations between pieces of evidence, to build knowledge, are all part of an information discovery process. This paper describes an approach to interactive topic analysis, as an information discovery conversation with a recommender system. We describe a model that motivates our approach, and an evaluation comparing interactive topic analysis with state-of-the-art topic analysis methods.

The skill to find and organize the right information has become paramount: searching and collecting information occupy a large portion of our daily productive time. While the focus rests on searching, major advances in numerous fields from artificial intelligence to information retrieval gear nowadays search engines. Search works much like a dictionary: one expresses information needs as a query and the engine responds with relevant results, where hopefully one finds the answer. Search is the ubiquitous access point to information. However, the goal is more often one of discovery. While search engines are more than information retrieval and, fitted with recommending approaches, use our browsing history to lead us to the expected results. The whole mechanism remains hidden and hardly supports the user in building relations and knowledge in itself (Marchionini 2006).

Even when a large repository of valuable resources is available, there is hardly a way to reorganize it along new topics. Collecting information about a new topic is rarely solved with a single query (Nolan 2008). It is rather an exploration and discovery process involving several queries intermingled with extensive reading of retrieved resources. Query terms are refined as results from previous searches are explored. Interesting results are collected along each step, but it is difficult to establish connections between them. The effort and time are mostly spent in careful reading and acquiring information from search results (often reading titles or text summaries) and in formulating new concepts to refine the search. Marchionini describes exploration in opposition to lookup search, involving both investigation and learning (Marchionini 2006). We extend the concept to a continuous process of information discovery entailing awareness, exploration, and explanation. Hence, the discovery process requires repetitive search and collecting pieces of information, formulating hypotheses and proving them.

One alternative is to rely on a different paradigm. Topic models use a generative approach (e.g., latent Dirichlet allocation) to capture the themes inherent to a document collection (Blei 2012). Topic modeling interfaces allow users to explore the collection contents at multiple levels, for example, using topic overview and zooming with keyword search (Ganguly et al. 2013). Topic models owe their flexibility to the fact that they do not correspond to any predefined taxonomy. In fact, the model generation process has no semantic information, but discovers patterns due to document-term co-occurrence. The flexibility turns into a caveat, as the topics generated are often not interesting or relevant to users (Hu et al. 2014). Topic models are costly to compute. The exploration and discovery process only works on a pre-existing collection. Although methods exist to interactively change a topic model (e.g., joining or splitting topics), they have been developed to improve the topic model rather than to support the exploration and discovery process. We seek a solution that allows the flexibility of topic models to explore a collection from the topics that run through it, but which also works on a continuously growing collection, which users can easily reorganize at will as their information needs change.

This paper puts forth a cognitive model of information discovery building on Marchionini's exploratory search notions (Marchionini 2006) and on information foraging theory (Pirolli 2007). It describes a novel prototype implementing the process using simple text mining techniques. Instead of trying to infer a hidden topic structure fully automatically, the proposed interactive approach works as a conversation between the user and a recommender system (RS) fostering the creation of a personalized theme structure. A visual analytics approach instantiates the information discovery model and enacts the conversation with the RS.

## Approach: Interactive Topic Analysis

Our approach focuses on building knowledge sustained in a cognitive model of exploratory search and discovery. The abstract model in Fig. 1 entails awareness, exploration and explanation, interlocked in a continuous discovery process.

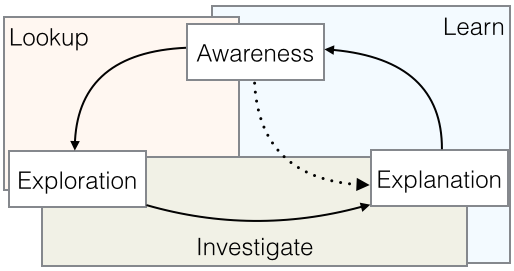


Figure 1: Information discovery model. A continuous process entailing awareness, exploration and explanation.

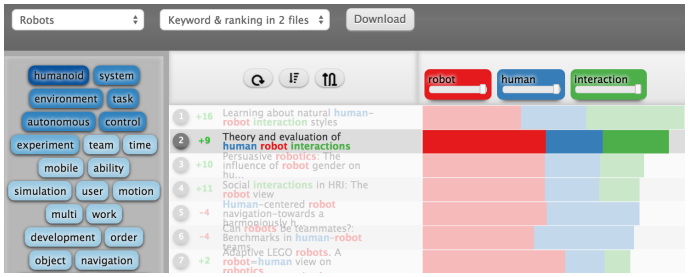


Figure 2: Visual analytics. Tags (left) summarizes content in keywords. Information needs are in the query box (top), which ranks the document set (bottom).

When the user becomes *aware* that she needs more evidence, she chooses a path and starts a focused *exploration*. While exploring, the user follows different leads, collects facts, and starts building hypotheses. As hypotheses mature, the user turns to test them, seeking *explanation* in the facts collected. Each of these stages contributes knowledge, as leads that the user becomes *aware* of, as facts acquired through *exploration*, or as relations elicited through *explanation*. In terms of information foraging (Pirolli 2007), awareness is guided by information scent which sustains the choice of leads to follow and patches to explore. In our case, patches are documents or sub-collections thereof. Foraging is the process of building the own topic structure along perceived relationships in the data collection. Our visual analytics interface instantiates this exploration and discovery model.

## Awareness

A great portion of the visual interface is designed to guide awareness. In particular, the topic summary is built from keywords extracted from the whole collection. To raise awareness of potential topics, keywords are presented in a box, organized and encoded in terms of their frequency, see Fig. 2. To this end, after preprocessing (e.g., singularizing, stemming), tf-idf (term frequency, inverse document frequency) is computed, and a global set of keywords is collected. They are sorted by the accumulated document frequency (DF). Global keywords summarize the collection contents, but to communicate how these contents are organized, the tag box interface makes the user aware of co-occurrences. On mouse over a keyword, two micro visualizations show a proportion of documents affected by choosing this keyword and a number of neighboring keywords, highlighted upon selection. Co-occurrence highlights help the user decide which keywords to choose for exploration.

## Exploration

Documents in the collection are initially shown as a list. Exploration is sustained upon a fast-ranking mechanism which ranks the recommendation list on-the-fly, as the user drags keywords to a query box to express her information needs. Each keyword can be assigned a weight using a slider, to reflect its importance in the query. The list turns into an interactive ranking, showing the relevance of documents to each change in the query box. A set of documents $D = d_1, ..., d_n$ is ranked from set of extracted keywords: $K = k_1, ..., k_m$ and a set of selected keywords: $T = t_1, ..., t_p, T \subseteq K$. The overall score for document $d_i$ is calculated as the sum of the weighted scores of its keywords matching selected keywords: $s_{d_i} = \sum_{j=1}^{p} w_{t_j} \times m_{l_i t_j}$, where $w_{t_j}$ is the weight assigned by the user to the selected keyword $t_j$, such that $\forall j : 0 \leq w_{t_j} \leq 1$; and $m_{d_i t_j}$ is the tf-idf score for keyword $t_j$ in document $d_i$.

$D$ is next sorted by overall score using the quicksort algorithm. Documents in $D$ are now elements of sequence $Q$ ordered by: $Q = (q_i)_{i=1}^{n}, q_i, q_{i+1} \in D \wedge s_{q_i} \geq s_{q_{i+1}}$.

Finally, the ranking position is calculated so that items with equivalent overall score share the same position. The position for a sorted document $q_i$ is calculated as

$$r_{q_i} = \begin{cases} 1 & \text{if } i = 0 \\ r_{q_{i-1}} & \text{if } s_{q_i} = s_{q_{i-1}} \\ r_{q_{i-1}} + |C| & \text{if } s_{q_i} \neg s_{q_{i-1}} \end{cases}$$

Where $C = q_j / s_{q_j} = s_{q_{j-1}}, 0 \leq j \leq i$ is the set of all the items with immediate superior overall score than $q_i$.

## Explanation

As the document set is ranked, a RankView shows the contribution each keyword has on the overall score of a document. It communicates visually the effects of adding and removing keywords to the query box. Additionally, when a document is selected, a text snippet or abstract is displayed in the focus view, highlighting selected keywords. Keyword colors and highlights are managed through a uniform color palette, allowing the user to quickly identify score contributions and text highlights. More importantly, keyword highlights drag the attention of the user to the portion of the text and its surroundings, promoting discovery of connected concepts.

# Evaluation

We designed a study to determine how our cognitive bottom-up approach elicits knowledge. Precisely, this preliminary evaluation motivates the construction of topic oriented document collections: participants had the task to form collections of items relevant to a suggested topic using our tool (U) or a recommendation list (L). The collected items were then used to compare the relationships established with our tool with state-of-the art topic analysis methods (LDA). The

experiment followed a repeated measures designs, with four iterations of the same tasks, varying the independent variables. Three tasks had to be performed per repetition: two focus exploration tasks (find five items most relevant to a set of keywords) and a broad exploration task (find five items relevant to a short text). For the focus exploration task, two or three keyword phrases were prepared for each iteration. They reflect the behavior of changing information interests in exploration, and finding more information about newly found topic. The broad exploration task reflects the need to clarify a broader textual description, building phrases to describe information needs.

## Data

Four collections were used, one per repetition, covering a range of technical, cultural, and scientific content: women in the workforce (WW), robotics (RO), augmented reality (AR), circular economy (CE). Each of these topics has a well-defined Wikipedia page that was fed to a federated system to build the collections. The federated system creates a query from the text of the page and forwards it to a number of content providers. It compiles a joint list with items from each provider, but it cannot establish how relevant the items are. The resulting list refers to the whole text with no indication of subtopics. Hence, subtopics for each task were chosen by the authors from the text in the Wikipedia page.

Having defined collections and subtopics, we created topic models from the collections as baseline state-of-the-art topic analysis. To generate the topic models, the *topicmodels* package of R was used, with variational expectation-maximization, using estimated $\alpha$. Finding the right parameters for topic modeling was a challenge. In the end, a topic model was created for each collection, with parameters ($K = 30$, $\alpha = estimated$) so that the focus exploration questions were covered by one or more topics. The broad exploration questions were adjusted, by choosing the highest number of words in the subject text appearing in a single topic. The document scores in each topic were modulated with the cumulative *tf_idf* scores for the chosen keywords.

## Procedure

Twenty four (24) participants took part in the study (*11* fem., *13* m., between 22 and 37 years old). They were recruited from the medical and computer science university population, and were not familiar with the topic areas selected for the study. The study had an introductory video showcasing the features of the system, followed by a short training session, to familiarize participants with the tool. When the participant was ready, the actual study started. For each condition, the system first showed a short text to introduce the topic. After reading the text, participants pressed start, opening the interface. At the beginning of each task, the items in the collection were ordered randomly, ensuring that an item would not appear in the same position again. The instructions for the task were shown in the upper part of the screen. In each task, participants had to collect five items. The task finished when the participant pressed the *finished* button. It was possible to finish without collecting all items. After finishing the three tasks, the test moved to the next condition.

Table 1: Emergent topics. Pearson product-moment correlations for top5 elements ranked (U), popular with ranking ($U_{MP}$), popular with list ($L_{MP}$), each compared to documents in selected topic from a topic model.

| | q1 | q2 | q3 |
|---|---|---|---|
| | WW | | |
| U | $r_{(11)} = -.00, p = .99$ | $r_{(6)} = .42, p = .29$ | $r_{(11)} = -.15, p = .60$ |
| $U_{MP}$ | $\mathbf{r_{(11)} = .68, p < .05}$ | $r_{(6)} = .39, p = .32$ | $\mathbf{r_{(11)} = -.65, p = .15}$ |
| $L_{MP}$ | $r_{(11)} = .16, p = .59$ | $r_{(6)} = -.55, p = .24$ | $\mathbf{r_{(11)} = .56, p < .05}$ |
| | RO | | |
| U | $r_{(11)} = -.41, p = .16$ | $r_{(9)} = .00, p = .98$ | $\mathbf{r_{(11)} = .55, p = 0.05}$ |
| $U_{MP}$ | $r_{(11)} = -.41, p = .16$ | $r_{(9)} = .05, p = .86$ | $r_{(11)} = .09, p = .74$ |
| $L_{MP}$ | $r_{(11)} = -.52, p = .06$ | $r_{(9)} = -.06, p = .8$ | $r_{(11)} = -.35, p = .23$ |
| | AR | | |
| U | $r_{(9)} = .38, p = .24$ | $r_{(9)} = .06, p = .84$ | $r_{(11)} = -.01, p = .95$ |
| $U_{MP}$ | $r_{(9)} = .08, p = .80$ | $r_{(9)} = .54, p = .08$ | $\mathbf{r_{(11)} = -.61, p < .05}$ |
| $L_{MP}$ | $r_{(9)} = .01, p = .96$ | $r_{(9)} = .00, p = .99$ | $\mathbf{r_{(11)} = -.58, p < .05}$ |
| | CE | | |
| U | $\mathbf{r - (6) = .82, p = .01}$ | $r_{(8)} = .23,$ | $\mathbf{r_{(11)} = -.57, p < .05}$ |
| $U_{MP}$ | $r_{(6)} = .64, p = .08$ | $r_{(8)} = .30, p = .39$ | $r_{(11)} = -.11, p = .7$ |
| $L_{MP}$ | $r_{(6)} = .53, p = .17$ | $r_{(8)} = .42, p = .22$ | $r_{(11)} = -.24, p = .4$ |

The order of task and collection were randomized with a balanced Latin Square. After each condition, participants had to fill a NASA TLX questionnaire to assess cognitive load, performance and effort.

## Results

The topics obtained with topic modeling (TM) were used as basis for comparison with the ranking (U). Additionally, we computed a list of most popular items collected with just the list ($L_{MP}$) and using the ranking ($U_{MP}$). To obtain normalized scores for the most popular lists we used frequency of choice (times chosen / trials, where trials=12). Figure 3 illustrates the overlap in score results. Topic models were very narrow. The ranking (U) tended to produce widespread results with low peaks when the requirements used many keywords. It seemed that there were common high scoring documents highlighted in the topic and also popular among participants. To establish if there is a valid correlation, we first collected the top five scoring items in each method (U, $U_{MP}$, $L_{MP}$, TM) and formed a matrix. The dimensions of the matrix varied due to the non perfect overlap (some methods had different top-scoring documents). We collected the scores from each method for all the documents in the matrix and computed *Pearson* product-moment correlations with TM: ($U - TM, U_{MP} - TM, L_{MP} - TM$). Table 1 summarizes the correlation results.

NASA TLX data were analyzed using a repeated measures ANOVA with independent variables tool, and dataset size. Post-hoc effects were computed using Bonferroni corrected pairwise comparisons. The two by two experimental design ensures that sphericity is necessarily met. A repeated measures ANOVA revealed a significant effect of tool on perceived workload F(1,23)=35.254, $p < 0.01, \epsilon = 0.18$. A Post-hoc paired-samples t-test revealed a significantly lower workload when using our tool ($p < 0.001$). Further, repeated measures ANOVA showed significant effects of tool in each dimension of the workload measure , as shown in Table 2.

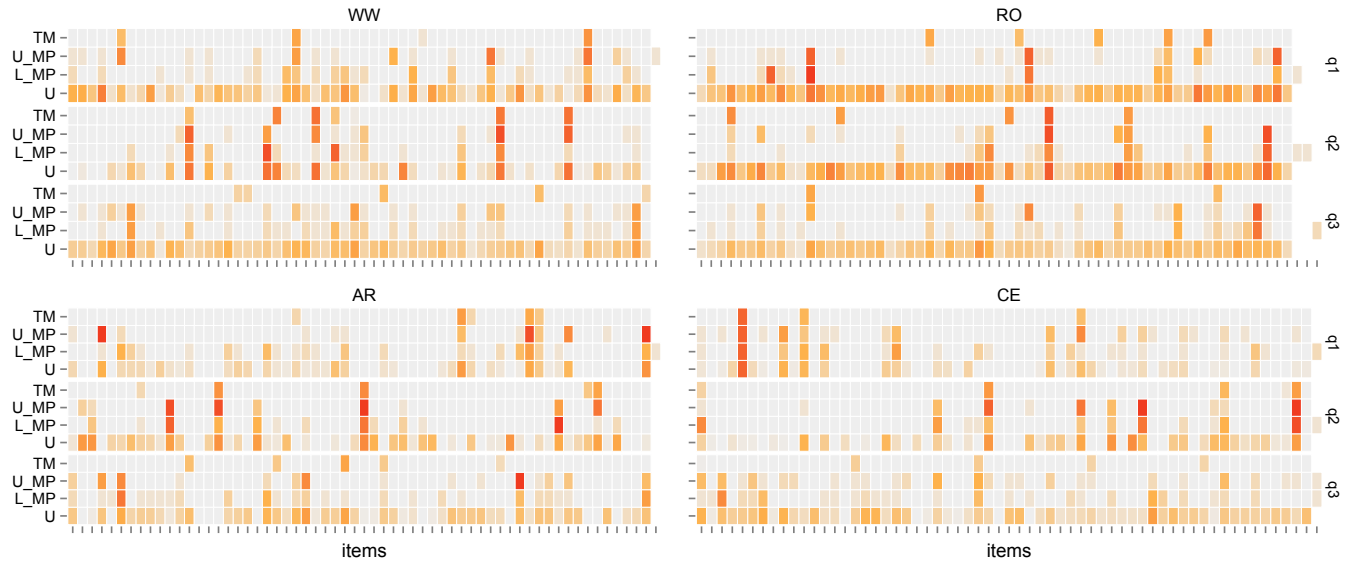


Figure 3: Document Scores. Documents found in the topic model (TM) also had high rank (U) and were picked often ($U_{MP}$, $L_{MP}$).

Table 2: Complexity: people found our tool incurs significantly lower workload in all dimensions

| Dimension | F(1,23) | p | $\epsilon$ |
|---|---|---|---|
| Mental Demand | 19.700 | $p < 0.05$ | 0.10 |
| Physical Demand | 14.520 | $p < 0.01$ | 0.07 |
| Temporal Demand | 7.720 | $p < 0.05$ | 0.05 |
| Performance | 11.800 | $p < 0.01$ | 0.10 |
| Effort | 48.600 | $p < 0.001$ | 0.22 |
| Frustration | 15.120 | $p < 0.01$ | 0.07 |
| Workload | 35.254 | $p < 0.01$ | 0.20 |

## Discussion and Outlook

Having found correlations with the results obtained with topic models was a surprise. Admittedly, topic models are used for exploration in a different way. Yet, this proves our goal: that cognitive topic analysis lead to comparable results. While being based on an established workflow of exploration and discovery, our approach lets the user experience the organization of the corpus along her information needs. Furthermore, socially emergent topics were surprisingly close in some cases to the topic model results. In the future we will continue to investigate the effect of social topic modeling and more tightly integrate RSs at different levels.

Attempting to recreate information needs along the division of topic models was a daunting task; it required several iterations of adjusting parameters, recreating the model, and searching through generated topics. Configurations that could answer one question, were not able to answer the next one, due to fragmented topics. One avenue of future research is the integration of topic modeling in the organization and recommendation of items. For instance, the keyword summary could be organized along known topics. Alternatively, the topics organization can be elicited through collaborative filtering from people's built collections. Finally, the study presented, while helping to validate the simplicity of our approach, lacks depth to thoroughly investigate all implications of our information discovery model. It does however validate that organizing the interaction in terms of the model simplifies the task, so users can perform with lower effort and frustration. In the future we will explore its applications in different tasks that require information discovery.

## Acknowledgments

This work was supported in part by CONICET. Know Center is supported by the Austrian Research Promotion Agency (FFG) under the COMET Program

## References

Blei, D. M. 2012. Probabilistic topic models. *Commun. ACM* 55(4):77–84.

Ganguly, D.; Ganguly, M.; Leveling, J.; and Jones, G. J. 2013. Topicvis: A gui for topic-based feedback and navigation. In *Proc. of the 36th Int. ACM SIGIR*, SIGIR '13, 1103–1104.

Hu, Y.; Boyd-Graber, J.; Satinoff, B.; and Smith, A. 2014. Interactive topic modeling. 95:423–469.

Marchionini, G. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49(4):41–46.

Nolan, M. 2008. Ia column: Exploring exploratory search. *Bulletin of the American Society for Information Science and Technology* 34(4):38–41.

Pirolli, P. 2007. Cognitive models of humaninformation interaction. In Durso, F., ed., *Handbook of Applied Cognition: 2nd Edition*. New York, NY, USA: Wiley & Sons, 2 edition.