

Exploratory Navigation and Selective Reading

Natwar Modani,^{1*} Paridhi Maheshwari,^{1*} Harsh Deshpande,^{2†}
 Saurab Sirpurkar,^{3†} Diviya,^{4†} Somak Aditya¹

¹Adobe Research, ²Indian Institute of Technology, Bombay

³Indian Institute of Technology, Madras, ⁴Indian Institute of Technology, Roorkee
 {nmodani, parimahe, saditya}@adobe.com, {hdeshpande5998, 13sau7, diviya7297}@gmail.com

Abstract

Navigating a collection of documents can be facilitated by obtaining a human-understandable concept hierarchy with links to the content. This is a non-trivial task for two reasons. First, defining concepts that are understandable by an average consumer and yet meaningful for a large variety of corpora is hard. Second, creating semantically meaningful yet intuitive hierarchical representation is hard, and can be task dependent. We present our system NAVIGATION.AI which automatically processes a document collection, induces a concept hierarchy using Wikipedia and presents an interactive interface that helps user navigate to individual paragraphs using concepts.

Introduction

When looking for information on a topic, one might not have a clear search goal and hence, a well-formed search query. Consider a US citizen who wants to understand the stand of the presidential candidates on various issues. Transcripts of the presidential debates are easily available, and the citizens can read it to form their opinion. However, there are multiple lengthy debates and understanding these requires significant effort from the reader. An alternative is to read a news media analysis, but it is not guaranteed to cover all aspects of the debate adequately and may be biased. Moreover, it may not be any shorter or easier to read/understand. Another example is of a student completing a research project. While they may find several relevant documents, understanding and organizing the information is a challenging task. In such situations, the ability to first understand the information content of documents (exploratory navigation), and then read the parts of deeper interest (selective reading) would be desirable.

We present our technology NAVIGATION.AI, that decomposes the documents into semantic units and assigns concise, human understandable concepts to each of them. We use Wikipedia article titles as concepts, and employ Gram-Schmidt orthonormalization to reduce redundancy amongst concepts and eliminate irrelevant ones. We then find a subset of labels which adequately represent the documents un-

*These authors contributed equally

†Work done while at Adobe

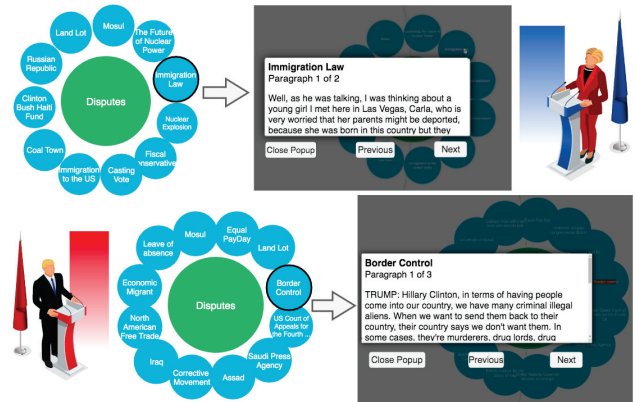


Figure 1: A subset of detected semantic concepts from 2016 presidential debate transcripts. Green nodes denote parent concepts and blue denotes leaves. Clicking on the node displays the associated paragraphs sequentially.

der consideration with a given number of labels. We leverage DBpedia (Auer et al. 2007), a structured knowledge base for Wikipedia, to find common ancestors of applicable labels to further reduce the required number of labels.

An interesting use-case for our framework is to explore the 2016 presidential debate transcripts. Looking at the tool’s output for Donald Trump’s statements in 2016 debates, the top topics in his mind seem to be: border control, jobs, tough foreign policies against ISIS-Syria etc. On the other hand, the most important topics for Hillary Clinton seem to be social security, jobs, election and women’s issues. We observe that the “dispute” concept (Figure 1) naturally groups many of the disputed topics that both leaders talk about. Navigating to the “dispute” bubble for Trump and Clinton, we observe that Trump talks about “Border Control” whereas Clinton discusses about “Immigration Law” and “Immigration to the US”. Reading the associated paragraphs reveals that Hillary Clinton is talking about families and children being separated from each other because of the legal structure, whereas Trump talks about sending back the criminals - posing contrasting stance on immigration.

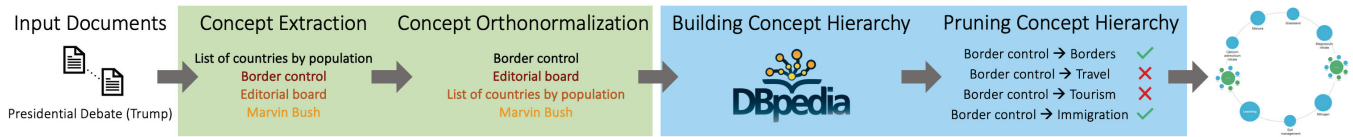


Figure 2: Our technique comprises of two key steps. The first step involves extracting a concise set of concepts for all text fragments with minimal semantic overlap. Next, we build a hierarchy to enable multiple abstraction levels in the concept space.

Related Work: For query-less search scenarios, multiple solutions have been proposed, such as automated Table-of-Contents (Erbs, Gurevych, and Zesch 2013) and topic analysis. While the former eases navigation for a single document, it does not address the problem of searching a corpus. Further, it suffers from abstract titles such as “introduction”, “motivation” which hardly convey any information about the section content. Alternatively, works on headline generation (Tan, Wan, and Xiao 2017) create condensed summaries for text but do not support navigation. Another common approach to thematically explore a corpus involves topic modelling techniques (Kim et al. 2016). Recent works include concept graph construction (Gordon et al. 2016) that support exploration through keywords and domain-specific concepts. All these methods detect latent topics, that are frequency distributions over words and are hard to interpret.

Technical Workflow and Results

There are two major aspects of our tool. First, finding human understandable and representative labels for each paragraph. Second, ensuring that a small number of labels can represent the document set adequately.

To obtain human understandable labels, we use Explicit Semantic Analysis (Gabrilovich and Markovitch 2009) to pick top-10 Wikipedia article titles as candidate labels. Next, to make the labels non-redundant (and therefore, more representative), we use Gram-Schmidt Orthonormalization (GSO). However, performing GSO for each paragraph independently will give us different labels for each. To ensure adequate coverage of the documents with a small number of labels, we start processing paragraphs which have more specific alignment with specific labels. The specificity is measured by normalizing the relevance scores of the labels, treating them as probabilities, and computing the entropy. Paragraphs are arranged in increasing order of entropy (more specific to less specific) to perform GSO. While computing the GSO for i^{th} most specific paragraph, we prefer the labels already used in first $(i - 1)$ most specific paragraphs. This ensures that if a previously used label is adequate, it will be reused instead of a new label. We pick top-5 labels for each paragraph based on their relevance scores after GSO.

$$u_k = v_k - \sum_{j=1}^{k-1} \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} u_j \quad (1)$$

Here, v_k is the k^{th} label being considered with relevance score as given by ESA and u_k is the corresponding GSO’ed relevance score for that label.

To pick a user specified number of labels that best represent the document set, we solve a constrained optimization

problem wherein the sum of contributions for a given number of labels is minimized. Further, we use DBpedia hierarchy to identify common ancestors instead of multiple descendant labels. As the number of levels between the ancestor label and given concept label increase, the representativeness of the ancestor decreases. We use an iterative greedy strategy to select labels. Once a label is selected to be included in hierarchy, we adjust the contributions of other labels to reflect their marginal contribution.

Results: We conducted a user-study to evaluate the quality of the extracted concepts. We curated 6 different datasets by picking the top results (3-5 documents) for the following web search queries: World War I, American History, Nitrate Leaching, Quality Assessment, The way we speak and Expanding Universe. Users are presented with an interactive interface comprising of the detected concepts, their hierarchy and paragraphs associated with them. Users then freely explore the visualization and learn about the document collection. Participants are asked to rate the quality of paragraphs associated with a concept and concepts for a paragraph on a 5-point Likert scale. The average ratings - 3.72 ± 0.77 and 3.56 ± 1.09 , indicate the goodness of the concept-paragraph mapping, with over 75% people rating it 4 or higher. Participants were further asked about the adequateness of the nodes in the visualization. 60% participants found the number of concepts to be just right to represent the documents.

References

- Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer. 722–735.
- Erbs, N.; Gurevych, I.; and Zesch, T. 2013. Hierarchy identification for automatically generating table-of-contents. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*.
- Gabrilovich, E., and Markovitch, S. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* 34:443–498.
- Gordon, J.; Zhu, L.; Galstyan, A.; Natarajan, P.; and Burns, G. 2016. Modeling concept dependencies in a scientific corpus. In *ACL (Volume 1: Long Papers)*, 866–875.
- Kim, M.; Kang, K.; Park, D.; Choo, J.; and Elmqvist, N. 2016. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE TVCG*.
- Tan, J.; Wan, X.; and Xiao, J. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, 4109–4115.