Supporting Research through WebMap – Multifaceted Semantic Link Induction in the Web

Mario M. Kubek
Georgia State University
Atlanta, GA, USA
mkubek@gsu.edu

Georg Philipp Roßrucker
FernUniversität in Hagen
Siegburg, Germany
g.rossrucker@gmail.com

Abstract

Carrying out research tasks is only inadequately supported, if not hindered, by current web search engines. This paper therefore proposes functional extensions of *WebMap*, a semantically induced overlay linking structure on the web to inherently facilitate research activities. These add-ons support the dynamic determination and regrouping of document clusters, as well as the creation of a semantic signpost.

1 Initial Situation

Web search engines exploit the explicit linking structure of the World Wide Web (WWW) to determine the relationships between web documents and assess the relevance and authority of content. Typically, hyperlinks are intentionally created and strategically placed by human efforts. However, it would be beneficial to also use semantically induced links between web documents and their content fragments to indicate topical relations and topically group potentially relevant web search results. This approach could facilitate labor-intensive research processes by automating the identification of relevant connections and topics.

2 WebMap's Architecture

To this end, in (Roßrucker, 2024), we introduced the *WebMap*, a novel solution to extending the existing linking structure of a hyperlinked network of text documents such as the WWW by a peer-to-peer-based semantic overlay, which induces and represents a distributed graph structure. The main idea is to embed a semantic and meaningful linking mechanism into the existing Web to make navigation and search independent of the existing – rather chronologically evolved – link structure. This is crucial because classical hyperlinks typically point to existing (older) content, putting new content at a disadvantage in terms of discoverability.

To achieve this, the global overlay linking structure is designed as a network of so called *Cluster Files*, generated and provided by the participating peers (web servers). Cluster files are identified by meaningful terms (text-representing centroids, TRCs (Kubek and Unger, 2016)) and consist of two sets of hyperlinks.

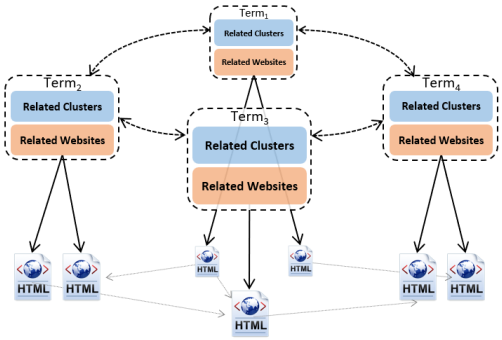

Figure 1: WebMap's Architecture

The first set refers to documents that are related within the given cluster's *context*, and the latter consists of links toward semantically related clusters, realizing traversable, bi-directional edges. Figure 1 illustrates how the overlaying linking structure of the *WebMap* extends the present linking structure of the underlying Web.

3 Improving WebMap

3.1 Neural Induction of Local Term Graphs

To derive the globally valid cluster assignment for documents and the necessary relations between clusters, peers of the *WebMap* make use of individual co-occurrence graphs (Jin and Srihari, 2007) that are induced by local text documents and capture simple syntagmatic term relations.

To obtain more meaningful document assignments and cluster associations in a harmonized manner, we propose to change the underlying mechanism from using co-occurrence graphs to local term graphs induced by large language models (LLMs) such as BERT (Devlin et al., 2019) and its variants. LLMs are able to convert input words into (sub)tokens and then into contextualized vec-

tor embeddings that capture the words' meanings in context.

To induce the required local term graphs, we propose to compare the word embeddings of meaningful elements such as nouns and proper nouns that appeared together in the local documents using cosine similarity. An edge is inserted between two elements if their cosine similarity is higher than a threshold value s, a hyperparameter to regulate the growth of the term graph. This way, the used LLM acts as a reference resource. We also would like to point out that the partaking peers do not necessarily have to agree upon a common LLM. On the contrary, depending on the individual web servers' content, domain-specific models such as SciBERT (Beltagy et al., 2019), and FinBERT (Araci, 2019) might be better suited to capture domain-specific characteristics and nuances.

3.2 Towards a Semantic Signpost

The assignment of individual documents to the global clusters is done by calculating their TRC terms using the local term graphs just discussed. Within the global cluster files, it is now possible to organize the documents based on their similarity as well as in relation to incoming search queries. Documents within a cluster will not only exhibit (flat) semantic similarities, but topical (hierarchical) dependencies as well. For instance, a document on the main topic *earthquake* could refer to contents that predominantly discuss its subtopics such as *seismic waves* and *movement of plate boundaries*.

The establishment of this intra-cluster linking structure yields a semantic signpost aiming to facilitate the targeted navigation to a topical direction of interest by lexically and semantically chaining documents. The topical dependencies can be uncovered by analyzing a cluster's document association graph and applying for instance a variant of the HITS algorithm (Kleinberg, 1999) on them. This way, a document's main (authorities) and source topics (hubs) can be identified. Table 1 illustrates this and shows that the main topic "Android" is greatly influenced by the subtopics "source code", "development", and "platform", which makes sense.

3.3 Handling Outliers and Reclustering

The described cluster assignment process works in a sequential manner and the number of cluster files does not need to be specified beforehand. However, this number is constrained by the finite set of terms in a natural language. Since sequential clustering is unsuitable to form compact clusters all the time, it is advisable to regularly run an iterative and density-based clustering algorithm such as (Komkhao et al., 2018) to identify meaningful and disjoint subclusters. Documents in subclusters with a low point density can be regarded as outliers and as candidates for re-clustering. The authors of this paper are in the process of analyzing this approach in more detail.

Term	Auth.	Term	Hub
Android	0.32	source code	0.19
Google	0.31	development	0.18
application	0.27	platform	0.14

Table 1: Terms with high authority and hub scores of the English Wikipedia article "Android" (operating system)

Limitations

While the proposed improvements of WebMap align cluster assignments with the use of LLMs, the approach is still limited to text documents on the Web. It is imperative that the employed LLMs fit well with the underlying text domains or at least constitute a *general* linguistic model. The semantic intra-cluster signposts are presently confined to a cluster's related documents, limiting the ability to link documents across different clusters. The improved WebMap still requires cooperative peers jointly providing the cluster files in common and good faith. Concerning implementation, adequate redundancy, fail-safeness and distribution of clusters among peers remain open tasks.

Ethics Statement

This research aligns with the ACL Code of Ethics[1], emphasizing fairness, transparency, and accountability. We prioritize user privacy, employ inclusive design practices, and ensure transparent algorithms. In data collection and analysis, we adhere to stringent privacy and consent standards. The proposed WebMap extensions aim to augment search and research capabilities responsibly, without causing harm. We actively welcome and engage in discussions on ethical considerations, demonstrating our commitment to addressing concerns transparently and promoting ethical practices in our work.

[1] https://www.aclweb.org/portal/content/acl-code-ethics

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

W. Jin and R. K. Srihari. 2007. Graph-based text representation and knowledge discovery. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, New York, NY, USA. ACM.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. In *J. ACM*, volume 46, pages 604—632, New York, NY, USA. ACM.

Maytiyanin Komkhao, Mario Kubek, and Wolfgang A. Halang. 2018. Sequential clustering and condensing the meaning of texts into centroid terms. In *Information Technology Journal*, volume 14, pages 1–10.

M. Kubek and H. Unger. 2016. Centroid terms as text representatives. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 99–102, New York, NY, USA. ACM.

Georg Roßrucker. 2024. *A Concept for a Distributed WebMap*, pages 37–67. Springer Cham.