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

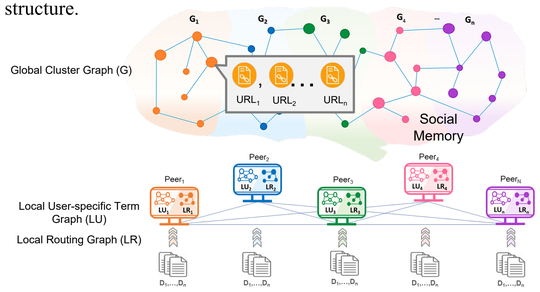
**Mario M. Kubek** and **Georg P. Roßrucker**
mkubek@gsu.edu g.rossrucker@gmail.com

## Initial Situation

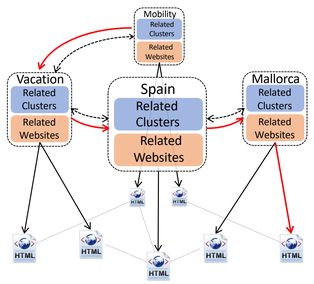
- Web search engines exploit the explicit linking structure of the World Wide Web (WWW) to determine the relationships between web documents and assess the relevance and authority of content. Yet, carrying out research tasks is only inadequately supported by current web search engines.
- Typically, hyperlinks 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.

## WebMap

- WebMap [1] is 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. The main idea is to embed semantic and meaningful links throughout the existing web, making navigation and search independent of the existing link structure.




- The global overlay linking structure is designed as a distributed 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 [2]) and consist of two sets of hyperlinks.

## LLM-induced Links

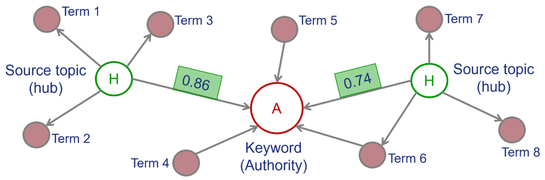
- Peers of the WebMap utilize individual cooccurrence graphs that capture simple syntagmatic term relations induced by local text documents to derive the globally valid cluster assignment for documents and establish the necessary relations between clusters.
- To obtain more meaningful document assignments and cluster associations in a harmonized manner that is commonly accepted and verifiable, we propose to change the underlying mechanism from using co-occurrence graphs to local term proximity graphs induced by Large Language Models (LLMs) [3] such as BERT and its variants, SciBERT, FinBERT, GPT-3, GPT-4, LLaMA-1, LLaMA-2, Mistral 7B, and others.

**Algorithm** Link induction and cluster assignment in the *WebMap*

```
<START>
Create local term proximity graph based on chosen LLM
Cluster assignment: for all local documents, derive TRCs (cluster identification)
    based on local term proximity graph
for all local documents do
    if respective cluster exists on the WebMap then
        Attach document link to cluster
    else
        Create cluster file locally and attach document link to cluster file
        Derive the shortest path from the set of existing cluster files represented in
        the local term proximity graph to the new cluster file
        Create cluster files and bi-directional links among them for all nodes on the
        path
    end if
end for
<END>
```

## A Semantic Signpost

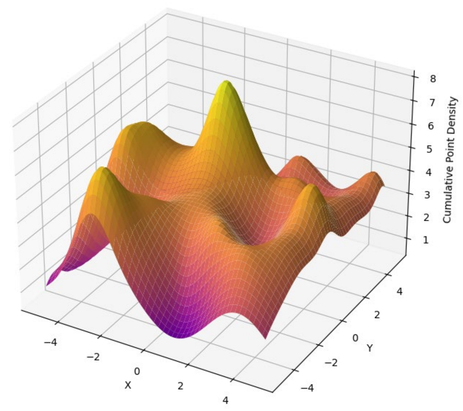
- Documents within a cluster will not only exhibit (flat, horizontal) semantic similarities, but topical (hierarchical) dependencies as well. For instance, a document on the main topic earthquake could refer to contents that predominantly discuss its important subtopics such as seismic waves and movement 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 mentioned topical dependencies can be uncovered by creating a cluster's directed and document-specific term associations graphs and applying for instance an extended variant [4] of the HITS algorithm.

## Subclusters

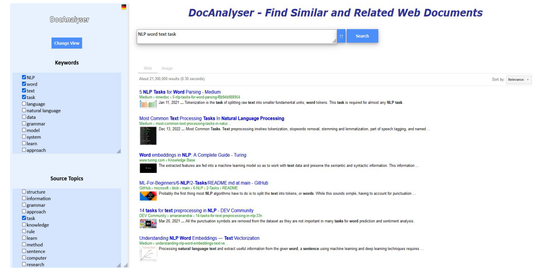
- Each cluster file is associated with a potentially large set of documents, which in turn can be associated with a set of subtopics and, in addition to the thematic dependencies described, can provide important clues for vertical navigation within a global cluster file. Therefore, it is advisable to regularly run an iterative and density-based clustering algorithm to identify those meaningful and disjoint subclusters.
- An approach to do so is based on a probabilistic interpretation, which considers feature vectors of items (here: the documents) as observations of a mixed population constituted by several overlapping populations, the sum of whose single unimodal distribution densities is a multimodal distribution density [5], which has several local maxima. Under the condition, that the single populations are sufficiently separated, it is assumed, that the local maxima characterize the regions in feature space where the single populations are concentrated, i.e. where clusters are expected.




- Those locations in feature space are searched, where a given data set exhibits local point concentrations with higher densities than in the respective vicinities.
- The search works by iteratively translating with a small step-size all feature vectors towards regions of higher point density [6]. By this process, the vectors gradually approach the local maxima. Merging into a single cluster all feature vectors thus arriving in the neighborhood of a certain location, an exhaustive and disjoint clustering of the data set is produced, with the number of these clusters derived from the characteristics of the data set, but not specified a priori.
- This also facilitates the detection of outliers by identifying clusters with low point density after the algorithm's execution. In the context of WebMap, documents in subclusters with a low point density can thus be regarded as outliers and as candidates for re-clustering. This could either mean that those documents will need to be assigned a different cluster file instead, or that they must be taken into account for a future subcluster assignment attempt.

## DocAnalyser.de

- WebMap may serve as a foundational search index for the development of novel search agents, especially those Artificial Intelligence (AI) tools aimed at supporting research processes. Example: Web service Docanalyser [7]




- AI tools targeting the research synthesis process of identifying, organizing, extracting, and combining studies can be suitably backed by WebMap's search infrastructure, too.

## Conclusion

- The proposed extensions to WebMap aim to support web-based research activities by leveraging advanced natural language processing techniques.
- By integrating LLMs and refining clustering algorithms, WebMap aims to provide users with more accurate, relevant, and comprehensive search results, ultimately enhancing their ability to navigate and explore complex information spaces on the web.
- WebMap has the potential to become a valuable tool for researchers, educators, and information seekers, facilitating seamless access to knowledge and insights across the vast expanse of the World Wide Web.

## Bibliography

1. Roßrucker, G.: A Concept for a Distributed WebMap. Springer Cham (2024)
2. Kubek, M., Unger, H.: Centroid terms as text representatives. In: Proceedings of the 2016 ACM Symposium on Document Engineering, pp. 99–102, ACM, New York, NY, USA (2016)
3. Minaee, S., Mikolov, T. et. al.: Large Language Models: A Survey. arXiv:2402.06196v2 [cs.CL] (2024)
4. Kubek, M.: Concepts and Methods for a Librarian of the Web. In: Studies in Big Data, Volume 62, Springer, Cham (2020)
5. Bock, H. H.: Automatische Klassifikation. Göttingen: Vandenhoeck & Ruprecht (1974)
6. Schnell, P.: Eine Methode zur Auffindung von Gruppen. In: Biometrische Zeitschrift, 6, pp. 47–48 (1964)
7. Website of DocAnalyser: https://www.docanalyser.de (2024)