University of Hagen @ TREC2021 News Track

Stefan Wagenpfeil*, Paul Mc Kevitt†, Matthias Hemmje*

*Faculty of Mathematics and Computer Science, University of Hagen, Germany

†Academy for International Science & Research (AISR), Derry/Londonderry, Ireland

*{firstname.lastname}@fernuni-hagen.de, †p.mckevitt@aisr.org.uk

Abstract

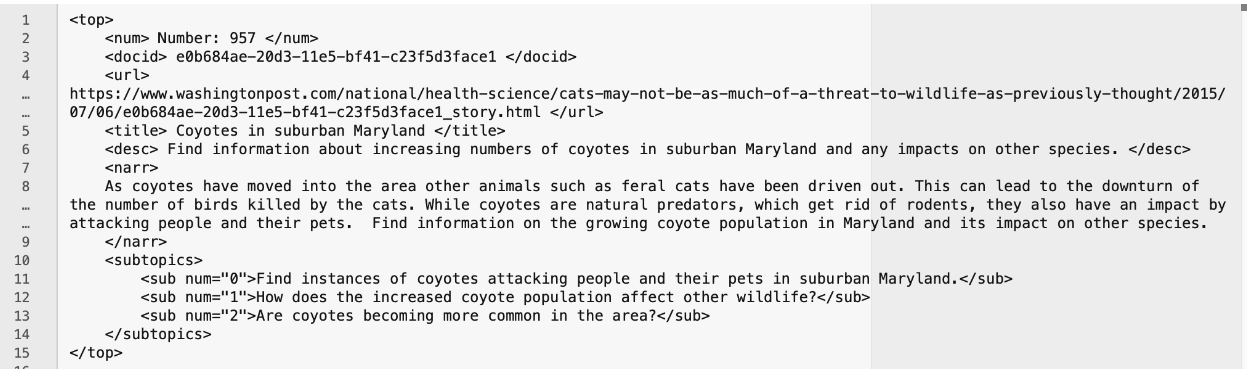
This paper describes discusses University of Hagen's approach for the TREC2021 News Track. The News Track aims at providing relevant background links to documents of the Washington Post article archive. Our submitted run is based on research and development in the field of multimedia information retrieval and employs a modified TFIDF (Text Frequency vs. Inverse Document Frequency) algorithm for topic modeling and matrix based indexing operations founded on Graph Codes for the calculation of similarity, relevance, and recommendations. This run was submitted as *FUH (Fernuniversität Hagen)* and obtained a nDCG@5 of 0.2655.

Index Terms

TFIDF; Graph Codes; similarity; relevance calculation; recommendation; feature graph; multimedia

I. Introduction, Motivation, and Document Collection

The research area of "Multimedia and Internet Applications" at the University of Hagen made inroads in the field of Multimedia Information Retrieval (MMIR) in recent decades. Some of these research results form an approach for the participation at the News Track of the TREC2021 conference. Typically, MMIR is aiming at the processing of multimedia content from various sources (e.g. images, audio, video, social media, and text). Relevant extracted information of these various multimedia objects are called "feature" [1]. The approach presented here is based on such multimedia features and can be applied to any kind of multimedia object including text. In the case of the TREC2021 News Track, it has been applied to the Washington Post article archive [2], which contains 728,626 news articles and blog posts from January 2012 through December 2020. The topic of the TREC2021 News Track is *Background Linking*, i.e. the identification of similar and/or relevant articles to a given reference article (called topic) and a set of corresponding questions (called subtopics). For this experiment, a set of 50 topics with additional 3-5 subtopics has been processed. An example of such a topic description is given in Figure 1. The articles of the Washington Post archive are stored in JSON format, and include fields for title, byline, date of publication, a section header, article text broken into paragraphs, and links to embedded images and multimedia (for the documents between the years 2012 and 2017).



```
<top>
    <num> Number: 957 </num>
    <docid> e0b684ae-20d3-11e5-bf41-c23f5d3face1 </docid>
    <url>
https://www.washingtonpost.com/national/health-science/cats-may-not-be-as-much-of-a-threat-to-wildlife-as-previously-thought/2015/
07/06/e0b684ae-20d3-11e5-bf41-c23f5d3face1_story.html </url>
    <title> Coyotes in suburban Maryland </title>
    <desc> Find information about increasing numbers of coyotes in suburban Maryland and any impacts on other species. </desc>
    <narr>
    As coyotes have moved into the area other animals such as feral cats have been driven out. This can lead to the downturn of
the number of birds killed by the cats. While coyotes are natural predators, which get rid of rodents, they also have an impact by
attacking people and their pets.  Find information on the growing coyote population in Maryland and its impact on other species.
    </narr>
    <subtopics>
        <sub num="0">Find instances of coyotes attacking people and their pets in suburban Maryland.</sub>
        <sub num="1">How does the increased coyote population affect other wildlife?</sub>
        <sub num="2">Are coyotes becoming more common in the area?</sub>
    </subtopics>
</top>
```

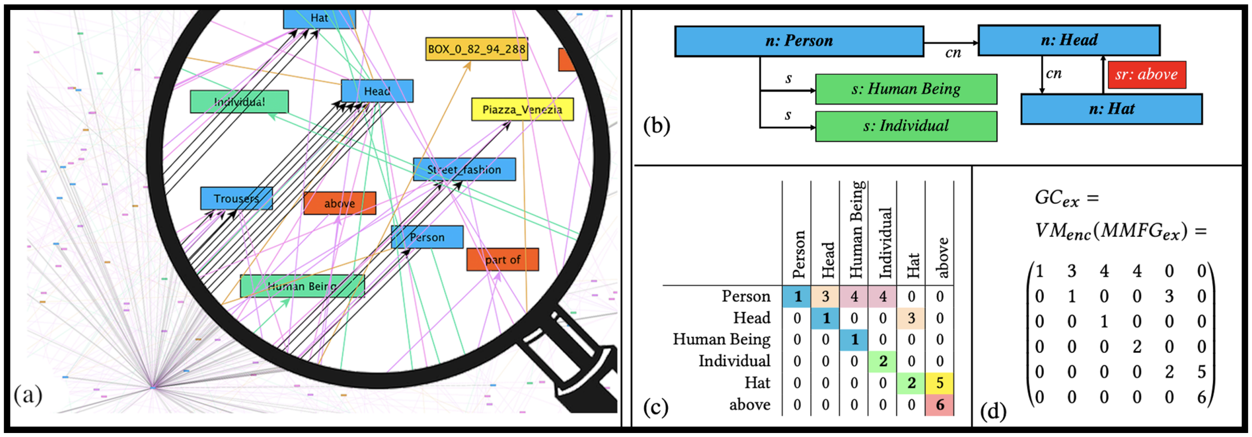
Fig. 1. Exemplary topic description.

In this paper, we discuss our approach and its application to text-only multimedia objects and discuss the results based on the TREC2021 News Track evaluation.

II. State of the art and related work

In previous related work [3][4][5], we introduced a *Generic Multimedia Analysis Framework (GMAF)*, which provides various plugins for multimedia feature detection in an extensible way. For the TREC2021 News Track, text-processing plugins are required, that are able to process the format of the Washington Post archive into a MultiMedia Feature Graph (MMFG). Such a MMFG is basically a graph structure and can be applied to the representation of MMIR features. A detailed description of this structure is given in [6]. Depending on the type of multimedia object, we discovered and showed [5][7], that graph algorithms have limited performance on large multimedia collections with a high level-of-detail [6]. Therefore, we introduced *Graph Codes* as a 2D projection of such graphs, which can utilize matrix calculations instead of graph traversal algorithms [8]. For these *Graph Codes*, we designed a special set of metrics to perform the parallelized calculation of similarity, recommendation, or inferencing operations. In [5] we showed, that the performance of these operations is in any case superior to graph traversal based operations. Until now, the main focus of the GMAF was the the detection of MMIR features in images and video. However, for partiipation in the TREC2021 News Track, various text based plugins have been developed. Therefore, we will also refer to text relevance algorithms, like TFIDF (Text Frequency vs.

Inverse Document Frequency) [9], named entity processing [10], and general concepts, like the Bag-Of-Words text-processing [11] in the remainder of this paper. Figure 2a shows an overview of such a MMFG, in the form of a conceptual visualisation. In Figure 2b, a simple example of MMIR features and their representation in the MMFG is given, which is then illustrated as a coloured (i.e. weighted) adjacency matrix in Figure 2c. Based on such a matrix, *Graph Codes* are calculated as shown in Figure 2d. This exemplary *Graph Code* GC_{ex} has been produced by the encoding function VM_{enc} (Valuation Matrix) applied to the example graph $MMFG_{ex}$.



	Person	Head	Human Being	Individual	Hat	above
Person	1	3	4	4	0	0
Head	0	1	0	0	3	0
Human Being	0	0	1	0	0	0
Individual	0	0	0	2	0	0
Hat	0	0	0	0	2	5
above	0	0	0	0	0	6

$$GC_{ex} = VM_{enc}(MMFG_{ex}) = \begin{pmatrix} 1 & 3 & 4 & 4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 3 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix}$$

Fig. 2. Overview of *Graph Codes* as: (a) visualisation (b) simple example, (c) corresponding matrix representation, (c) final *Graph Code*

III. Modeling and Algorithm Overview

Our goal for TREC2021 is to demonstrate, that the concept of *Graph Codes* not only improves the efficiency of MMIR, which has already been evaluated and outlined in [8]. We furthermore want to demonstrate, that *Graph Codes* deliver high effectiveness and accurate query results. Therefore, we chose to employ existing text processing algorithms and evaluate, whether *Graph Codes* based on these algorithms deliver high *efficiency* and *effectiveness* without any loss of precision. For the TREC2021 task, we chose the TFIDF algorithm as a purely statistical approach for the modeling of the text processing, and integrated it into the GMAF for MMIR. It may be noted, that the application of any other text processing algorithm would also be supported by the GMAF. Figure 3 shows the overview of our approach, which can be detaluled further as follows:

1) apply a TFIDF calculation to the documents and terms in the Washington Post archive. This task is performed to detect relevant and irrelevant words. Irrelevant words are then removed, relevant words are stored in a relevance database, from where they can be further used. This processing

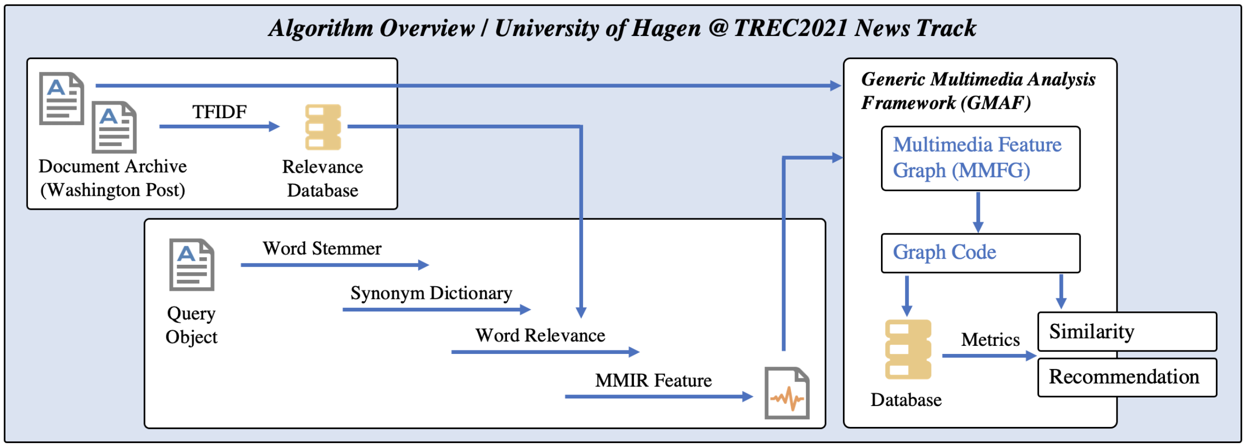


Fig. 3. Algorithm overview of the processing steps for the TREC2021 News Track.

can be applied to the complete collection of documents in advance. Whenever new documents are added to the collection, a re-processing can be required. However, due to the number of documents, this re-processing does not have to be performed immediately when a single document is added, but can be calculated as a background task with reduced priority.

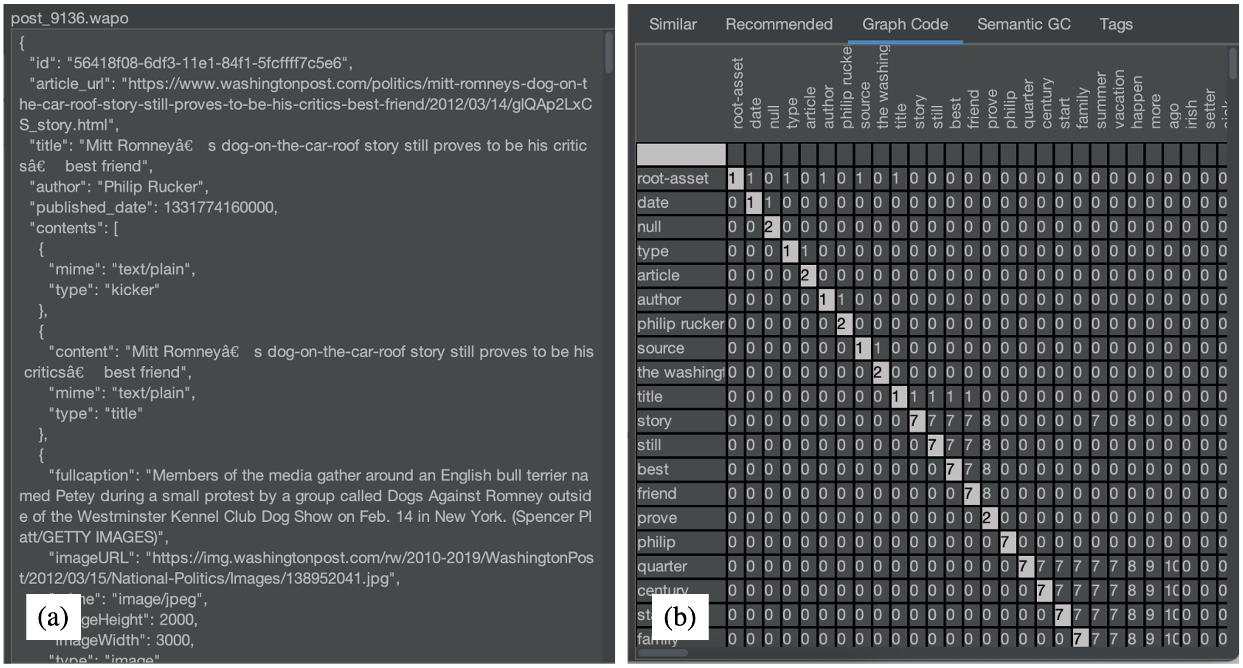
2) for each document in the collection, a *Graph Code* is calculated and stored in a MMIR database.
3) for each topic of the task description (i.e. the query documents), we apply:
   - a Word Stemmer, that reduces the detected term of each word to its grammatical base form
   - a query to a synonym dictionary, which returns all known synonyms for a given term
   - a check, if the word (or its synonyms) are relevant within the overall collection
   - a MMIR feature extraction, that is based on a Bag-Of-Words algorithm and calculates relevant words and word-stems of a given article's content

   For both the Word Stemmer and the synonym dictionary, the *Oxford English Dictionary* has been chosen, which is available for download and scientific use [12].
4) the detected features of the topic are then transferred to the GMAF, where they are converted into *Graph Codes*. Here, the *Graph Code* metrics for similarity and recommendations are applied to order the collection as a ranked list (see also [8]).

For the modeling of the subtasks experiment, we append the text representing the subtask to the original query text, so that also the detected terms of each subtask's question will be indexed. This ensures, that both question and original article text are considered in the algorithm. Further details on the implementation will be given in the next section.

IV. Implementation and Evaluation

The algorithm has been implemented in Java, based on the GMAF, and follows the approach discussed in the previous section. The TFIDF calculation in our setup returned a global word count of 41,920 unique words in 728,627 files. 34,113 words have been marked as relevant, 18,618 as irrelevant. After feature detection, the resulting MMFGs and *Graph Codes* had an average of 224 relevant keywords per article, that are employed for similarity and recommendation calculations. The *Graph Code* encoding has been modified in a way, that the terms of a sentence are related via relationship attributes. If a term occurs in several sentences, the corresponding *Graph Code* fields are also filled accordingly, leading to an increase in the relevance of this term. The algorithms for similarity and the detection of recommendations remain unchanged (see [8]). An example for such a generated *Graph Code* is shown in Figure 4.



post_9136.wapo

```
{
  "id": "56418f08-6df3-11e1-84f1-5fcffff7c5e6",
  "article_url": "https://www.washingtonpost.com/politics/mitt-romneys-dog-on-t
he-car-roof-story-still-proves-to-be-his-critics-best-friend/2012/03/14/gIQAp2LxC
S_story.html",
  "title": "Mitt Romneyâ€   s dog-on-the-car-roof story still proves to be his critic
sâ€    best friend",
  "author": "Philip Rucker",
  "published_date": 1331774160000,
  "contents": [
    {
      "mime": "text/plain",
      "type": "kicker"
    },
    {
      "content": "Mitt Romneyâ€   s dog-on-the-car-roof story still proves to be his
 criticsâ€    best friend",
      "mime": "text/plain",
      "type": "title"
    },
    {
      "fullcaption": "Members of the media gather around an English bull terrier na
med Petey during a small protest by a group called Dogs Against Romney outsid
e of the Westminster Kennel Club Dog Show on Feb. 14 in New York. (Spencer Pl
att/GETTY IMAGES)",
      "imageURL": "https://img.washingtonpost.com/rw/2010-2019/WashingtonPos
t/2012/03/15/National-Politics/Images/138952041.jpg",
      "mime": "image/jpeg",
      "imageHeight": 2000,
      "imageWidth": 3000,
      "type": "image"
```

	root-asset	date	null	type	article	author	philip rucke	source	the washing	title	story	still	best	friend	prove	philip	quarter	century	start	family	summer	vacation	happen	more	ago	irish	setter	
root-asset	1	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
date	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
null	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
type	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
article	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
author	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
philip rucker	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
source	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
the washingt	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
title	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
story	0	0	0	0	0	0	0	0	0	0	7	7	7	7	8	0	0	0	0	0	7	0	8	0	0	0	0	0
still	0	0	0	0	0	0	0	0	0	0	0	7	7	7	8	0	0	0	0	0	0	0	0	0	0	0	0	0
best	0	0	0	0	0	0	0	0	0	0	0	0	7	7	8	0	0	0	0	0	0	0	0	0	0	0	0	0
friend	0	0	0	0	0	0	0	0	0	0	0	0	0	7	8	0	0	0	0	0	0	0	0	0	0	0	0	0
prove	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
philip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
quarter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	7	7	7	7	7	8	9	10	0	0	0
century	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	7	7	7	7	8	9	10	0	0	0
st	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	7	7	7	8	9	10	0	0	0
family	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	7	7	8	9	10	0	0	0

Fig. 4. Exemplary *Graph Code* calculated from a random selected element of the Washington Post archive. (a) source article, (b) excerpt of the calculated *Graph Code*.

The evaluation of our algorithm is performed according to the TREC2021 News Track standards, by employing the corresponding tools, and is based on the TREC run results provided, which have been produced by the TREC2021 team in an anonymized evaluation. Table I summarizes these results and shows the overall average of all topics, including a selection of three individual topics[1]. The columns

[1] In the result list for our experiment, there were zero results for 13 topics. These topics have been removed from the result list in the table (topic = all)

method	*topic*	*5%*	*10%*	*15%*	*20%*	*min*	*mean*	*max*
Precision	all	0.4768	0.4291	0.3727	0.3285			
Recall	all	0.6229	0.1036	0.1301	0.1491			
nDCG@5	all	0.2655	0.2731	0.2674	0.2664	0.0	0.3148	0.6342
Precision	960	0.8000	0.7000	0.5333	0.5000			
Recall	960	0.1143	0.2000	0.2286	0.2857			
nDCG@5	960	0.3681	0.4558	0.4163	0.4290	0.0	0.278	0.552
Precision	947	0.8000	0.8000	0.6667	0.5000			
Recall	947	0.0870	0.1739	0.2174	0.2174			
nDCG@5	947	0.4353	0.4703	0.5293	0.4995	0.0	0.371	0.890
Precision	937	0.2000	0.1000	0.1333	0.1000			
Recall	937	0.0213	0.0213	0.0426	0.0426			
nDCG@5	937	0.0424	0.0323	0.0618	0.0539	0.0	0.152	0.761

TABLE I

ALGORITHM EVALUATION SELECTED EXPERIMENTAL RESULTS.

”min”, ”mean”, and ”max” are taken from the official aggregated results of all participants. The results indicate, that our experiment produced values for precision and recall, that are below the aggregated average of all participants. An analysis of the individual topic runs shows, that there are several topics, which have been processed very well (e.g. topic number 960 and 947), whilest others (e.g., number 937) produced low values for precision and recall. We discovered, that the reason for this is the threshold of the TFIDF algorithm, which marked several search terms of the topics as being irrelevant and hence failed to produce accurate results. For example, topic 937 is described with ”What is the Middle East Respiratory Syndrome (MERS)”. However, the TFIDF algorithm marked almost any word in this sentence as irrelevant, as the terms ”what”, ”is”, ”the” are irrelevant because they exist in almost any other document, and the words ”Respiratory”, and ”Syndrome” are marked irrelevant, because they appear in almost no other document. Therefore, the overall result of this topic is below the average.

In general, for these kind of very specific topic descriptions, the TFIDF algorithm is not the best choice and should be replaced or reconfigured with other parameters. However, the results show, that the introduction of *Graph Codes* does not influence effectiveness of MMIR in a negative way, as the results of runs, without these mentioned TFIDF flaws are above the average of the overall evaluation. This demonstrates, that MMIR based *Graph Codes* are both efficient and effective.

V. Conclusion and Summary

As discussed in the previous section, improvements to our approach have to be made to achieve an improved effectiveness. In particular, the TFIDF algorithm and its current settings filters too much data. Hence, further runs with different parameters for the TFIDF threshold have been performed, which increase the values for precision and recall up to 15%. These results will reported in future work and not here, as they have not been obtained during the TREC2021 News Track. Furthermore, our experiment shows, that a pure statistical algorithm can produce results comparable with the average of all other TREC2021 participants. With incorporation of the proposed improvements, the results are better than average. As our algorithm is purely based on MMIR features, images, audio, or video content could also be processed similarly. In particular, the fusion of these various multimedia feature sources will further increase the effectiveness of MMIR. This is subject of ongoing and future work.

References

[1] M. Nixon, *Feature Extraction and Image Processing for Computer Vision*. 125 London Wall London EC2Y 5AS UK: Academic Press by Elsevier Ltd., 2020.

[2] T. W. Post. (Mar. 2021). “Washington post archives”, [Online]. Available: https://www.washingtonpost.com.

[3] S. Wagenpfeil and M. Hemmje, *Towards ai-bases semantic multimedia indexing and retrieval for social media on smartphones*, SMAP 2020 Conference, Sep. 2020.

[4] S. Wagenpfeil, F. Engel, P. McKevitt, and M. Hemmje, *Semantic query construction and result representation based on graph codes*, BIRDS 2021: Bridging the Gap between Information Science, Information Retrieval and Data Science, Mar. 2021. [Online]. Available: http://ceur-ws.org/Vol-2863/#paper-06.

[5] S. Wagenpfeil, P. McKevitt, and M. Hemmje, “Graph codes - 2d projections of multimedia feature graphs for fast and effective retrieval”, Mar. 2021.

[6] S. Wagenpfeil, F. Engel, P. M. Kevitt, and M. Hemmje, “Ai-based semantic multimedia indexing and retrieval for social media on smartphones”, *Information*, vol. 12, no. 1, 2021, ISSN: 2078-2489. DOI: 10.3390/info12010043. [Online]. Available: https://www.mdpi.com/2078-2489/12/1/43.

[7] M. Needham, *Graph Algorithms*. 1005 Gravenstein Highway North Sebastopol CA 95472: O’Reilly Media Inc., 2019, ISBN: 978-1-492-05781-9.

[8] S. Wagenpfeil, B. Vu, P. Mc Kevitt, and M. Hemmje, “Fast and effective retrieval for large multimedia collections”, *Big Data and Cognitive Computing*, vol. 5, no. 3, 2021, ISSN: 2504-2289. DOI: 10.3390/bdcc5030033. [Online]. Available: https://www.mdpi.com/2504-2289/5/3/33.

[9] T. Eljasik-Swoboda, F. Engel, and M. Hemmje, “Explainable and transferrable text categorization”, in *Data Management Technologies and Applications*, S. Hammoudi, C. Quix, and J. Bernardino, Eds., Cham: Springer International Publishing, 2020, pp. 1–22, ISBN: 978-3-030-54595-6.

[10] C. Nawroth, F. Engel, T. Eljasik-Swoboda, and M. Hemmje, “Towards Enabling Emerging Named Entity Recognition as a Clinical Information and Argumentation Support”, in *Proceedings of the 7th International Conference on Data Science, Technology and Applications, DATA 2018*, SciTePress, 2018, pp. 47–55, ISBN: 978-989-758-318-6. DOI: 10.5220/0006853200470055.

[11] J. Leveling, “Interpretation of coordinations, compound generation, and result fusion for query variants”, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’13, Dublin, Ireland: Association for Computing Machinery, 2013, 805–808, ISBN: 9781450320344. DOI: 10.1145/2484028.2484115. [Online]. Available: https://doi.org/10.1145/2484028.2484115.

[12] (Aug. 2021). “The oxford english dictionary”, Oxford University Press, [Online]. Available: https://www.oed.com, Download: 11.08.2021.