Generating Semantic Aspects for Queries

Dhruv Gupta
Max Planck Institute for Informatics and Graduate School of Computer Science
Saarbrücken, Germany
dhgupta@mpi-inf.mpg.de

Klaus Berberich
Max Planck Institute for Informatics and htw saar
Saarbrücken, Germany
kberberi@mpi-inf.mpg.de

Jannik Strötgen
Max Planck Institute for Informatics
Saarbrücken, Germany
jstroetge@mpi-inf.mpg.de

Demetrios Zeinalipour-Yazti
University of Cyprus
Cyprus
dzeinali@mpi-inf.mpg.de

ABSTRACT

We present an approach to explore news archives by automatically generating semantic aspects for their navigation. Given a keyword query as an input, we utilize semantic annotations present in the pseudo-relevant set of documents for generating the aspects. Our approach to generate the aspects considers the salience of the annotations by modeling their semantics as well as considering their co-occurrence in the pseudo-relevant set of documents. The generated aspects are also beneficial for representing documents in a structured manner. We show preliminary results on two news archives demonstrating the quality of the generated aspects over a testbed of more than 5,000 aspects derived from Wikipedia.

ACM Reference Format:
Dhruv Gupta, Klaus Berberich, Jannik Strötgen, and Demetrios Zeinalipour-Yazti. 2018. Generating Semantic Aspects for Queries. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203900

1 INTRODUCTION

To explore news archives using only keywords offers us limited opportunities for their meaningful exploration. Nowadays, commercial search engines provide features such as *knowledge panels* [8] for queries regarding popular named entities as effective means of exploring web search results. This functionality, however, is not offered for ad-hoc keyword queries and is limited to those queries that concern popular named entities or selected search terms.

Also, users often struggle to convey their information needs clearly. A web query log analysis showed that ca. 46% of users performed query reformulation to better reflect their information needs [6]. To assist users in finding relevant documents quickly, we propose to provide semantic aspects as means of navigation. We generate these by utilizing semantic annotations present in the form of temporal expressions, disambiguated geographical locations, and other named entities in document contents.

In short, we generate aspects to informational queries as follows. First, we retrieve pseudo-relevant documents that mention the keywords present in the user's query. Second, we analyze semantic annotations in the form of temporal expressions, disambiguated geographical locations, and other named entities by considering their salience in models informed of their inherent semantics. Third, we analyze the annotations sharing different semantics by considering their co-occurrence in the set of pseudo-relevant documents.

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203900

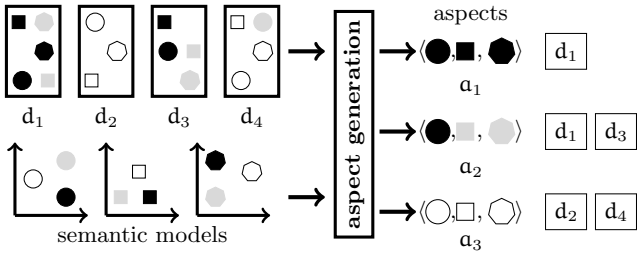

Figure 1: We generate aspects for a given set of documents by taking into account the salience of the annotations in models that are informed of their inherent semantics. We consider co-occurrence of annotations sharing different semantics. Generated aspects are also useful to provide a structured representation of documents (e.g., d_1 can be modeled as containing a_1 and a_2).

2 APPROACH

We now describe in brief the approach to generate the semantic aspects. Figure 1 gives an overall schema of our proposed approach.

Consider a news archive with documents $\mathcal{D} = \{d_1 \dots d_N\}$. Given a keyword query q, we use a retrieval method (e.g., Okapi BM25) to obtain a set of pseudo-relevant documents $\mathcal{R} \subset \mathcal{D}$ containing the query terms. Each document $d \in \mathcal{R}$ is enriched with annotations in the form of temporal expressions, disambiguated locations, and disambiguated persons and organizations. We can therefore represent a document using these annotations and the words it contains as:

$$d = \{d_{\mathcal{W}}, d_{\mathcal{T}}, d_{\mathcal{G}}, d_{\mathcal{E}}\}. \quad (1)$$

In Equation 1, the document d is modeled as a bag-of-words $d_{\mathcal{W}}$, bag-of-temporal-expressions $d_{\mathcal{T}}$, bag-of-locations $d_{\mathcal{G}}$, and bag-of-other-named-entities $d_{\mathcal{E}}$.

Generating Semantic Aspects. An aspect represents the salience of the annotations in models that are informed of their semantics as well as the co-occurrence of annotations sharing different semantics. An aspect a consists of factors x that exhibit the aforementioned properties. Concretely, an aspect is represented as:

$$a = \langle x_{\mathcal{T}}, x_{\mathcal{G}}, x_{\mathcal{E}} \rangle. \quad (2)$$

A generated sample aspect is shown in Figure 2. In Equation 2, $x_{\mathcal{T}}$ represents a time interval of interest. We leverage the work by Gupta and Berberich [5] to generate them. In short, time intervals of interest are obtained by computing overlaps of temporal expressions (i.e., expressions that convey mentions of time that are explicit, implicit, and relative) in a time model informed of temporal

⟨ Time: [2016,2016], Entities:[WIKI:RYAN_LOCHTE][WIKI:MICHAEL_PHELPS] [WIKI:MISSY_FRANKLIN][WIKI:CONOR_DWYER][WIKI:KATIE_LEDECKY] [WIKI:ALY_RAISMAN][WIKI:SIMONE_BILES][WIKI:NATHAN_ADRIAN] [WIKI:ALEXANDER_MASSIALAS][WIKI:ANTHONY_ERVIN] [WIKI:GABBY_DOUGLAS][WIKI:SUN_YANG],Locations:[WIKI:UNITED_STATES] [WIKI:CALIFORNIA][WIKI:NEW_YORK_CITY][WIKI:LOS_ANGELES] ⟩

Figure 2: A sample generated semantic aspect for the query OLYMPIC MEDALISTS. In the aspect several famous swimmers and gymnasts are shown. The time interval signifies the 2016 Summer Olympics. While the location signifies the nationality of the athletes.

uncertainty. Furthermore, each temporal expression is weighted by the score of the document (i.e., the score returned by the retrieval method) containing the temporal expression.

In Equation 2, the factors $\mathrm{x}_{\mathcal{G}}$ and $\mathrm{x}_{\mathcal{E}}$ correspond to interesting locations and other named entities, respectively. These factors are identified by modeling them using their Wikipedia page links. A person, organization, or location is deemed interesting if it is highly related to other named entities present by using the Jaccard similarity. We again weight their relatedness with the document's score. We thereby give more importance to those locations and other entity factors that are present in highly relevant documents.

Finally, we determine whether the various factors having different semantics in Equation 2 co-occur frequently in some subset of the pseudo-relevant document set $\mathcal{R}$. We determine this by checking that each factor in the aspect is salient above a given threshold and is generated from the same subset of $\mathcal{R}$.

3 EVALUATION

Semantically Annotated News Archives. We consider two news archives for evaluation: (i) the New York Times Annotated Corpus, which contains news articles published during the years 1987-2007 [2]. (ii) STICS, which contains news articles crawled from the Web between the time period 2013-2016 [9]. Semantic annotations for documents were obtained by using two types of natural language processing tools. First, we obtained temporal expressions for the documents using the HeidelTime temporal tagger [12]. It provides us with resolved time intervals for explicit, implicit, and relative temporal expressions. Second, we obtained, disambiguated named entities for the documents by using AIDA as the named entity recognition and disambiguation tool [10]. By using AIDA, we are able to resolve the disambiguated locations and other named entities to their Wikipedia pages. Table 1 shows the statistics regarding the news archives and the annotations for the documents.

Testbed. In order to evaluate the quality of the generated aspects, we turned to Wikipedia. Specifically, we created a testbed of 5,122 aspects by exploiting various event tables present in Wikipedia. We selected the queries and their corresponding pages by using the Wikipedia's page on "List of lists of lists" [1]. We parsed these Wikipedia pages and additionally used external resources, such as [4], to obtain the ground-truth aspects. The query keywords and their aspects counts are shown in Table 2.

Table 1: Statistics for the news archives and contained annotations for up to 10,000 documents obtained for each evaluation query.

COLLECTION	$\mathfrak{n}_{\text{documents}}$	μ_{time}	μ_{location}	μ_{entity}
NEW YORK TIMES	1,679,374	12.50	8.65	16.25
STICS	4,075,720	10.09	5.93	10.89

Table 2: Queries with aspect counts in brackets.

Achievements [1,508]: nobel prize [114] | olympic medalists [48] | oscars [1, 167] | paralympic medalists [24] | space shuttle missions [155]
Disasters [1,536]: aircraft accidents [513] | avalanches [56] | earthquakes [39] | epidemics [211] | famines [133] | genocides [35] | hailstorms [39] | landslides [85] | nuclear accidents [26] | oil spills [140] | tsunamis [88] | volcanic eruptions [171]
Politics [2,078]: assassinations [130] | cold war [81] | corporate scandals [44] | proxy wars [34] | united states presidential elections [57] | terror attacks [316] | treaties [1, 057] | wars [359]

Results. We discuss initial results that measure the precision and recall of our proposed approach against a naïve baseline considering all documents returned by Okapi BM25 with disjunction operator. For the baseline, we assume each document is equivalent to an aspect (i.e., Equation 1 and 2 are identical by disregarding the bag-of-words). The preliminary results in Table 3 show that for both the news archives our approach provides both high precision and recall values than the baseline which can only outperform our method in terms of recall as it considers all the documents retrieved for the query. The smaller average size of the aspects generated ($|\mathcal{A}|$) by our method additionally reflects that many documents share similar structure that can be exploited for meaningful navigation.

Table 3: Precision and recall results for news archives.

	NEW YORK TIMES			STICS		
	AVG. $\lvert\mathcal{A}\rvert$	PRECISION	RECALL	AVG. $\lvert\mathcal{A}\rvert$	PRECISION	RECALL
OKAPI BM25	3,379	0.098	**0.152**	3,796	0.072	0.113
OUR METHOD	1,638	**0.264**	0.148	480	**0.229**	**0.134**

4 RELATED WORK

An important approach in discovering structure within text documents was given by Hearst and Plaunt [7]. Their TextTiling algorithm structures documents with sub-topics identified using only text. Koutrika et al. [11] furthermore investigated how topics contained in documents can be helpful in information consumption by generating reading orders. From the perspective of faceted search, Ben-Ytizhak et al. [3] presented methods to identify facets for document exploration. Our approach in contrast has looked at modeling the semantics underlying the annotations when measuring salience. We also take into account co-occurrence of annotations sharing different semantics when generating aspects for archive exploration.

REFERENCES

[1] List of lists of lists. en.wikipedia.org/wiki/List_of_lists_of_lists
[2] The NYT Annotated Corpus. catalog.ldc.upenn.edu/LDC2008T19
[3] Ben-Yitzhak O. et al. Beyond basic faceted search. In *WSDM'08.*
[4] Bhagavatula C. S. et al. *TabEL: Entity Linking in Web Tables.* In *ISWC'15.*
[5] Gupta D. and Berberich K. Identifying Time Intervals of Interest to Queries. In *CIKM'14.*
[6] Hearst M. A. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press, New York, NY, USA.
[7] Hearst M. A. and Plaunt C. Subtopic Structuring for Full-Length Document Access. In *SIGIR'93.*
[8] Henry J.W. Providing Knowledge Panels With Search Results. (May 2 2013). US Patent App. 13/566,489.
[9] Hoffart J. et al. STICS: searching with strings, things, and cats. In *SIGIR'14.*
[10] Hoffart J. et al. Robust Disambiguation of Named Entities in Text. In *EMNLP '11.*
[11] Koutrika G. et al. Generating reading orders over document collections. In *ICDE'15.*
[12] Strötgen J. and Gertz M. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298.