Event-Centered Information Retrieval Using Kernels on Event Graphs

Goran Glavaš and **Jan Šnajder**
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{goran.glavas,jan.snajder}@fer.hr

Abstract

Traditional information retrieval models assume keyword-based queries and use unstructured document representations. There is an abundance of event-centered texts (e.g., breaking news) and event-oriented information needs that often involve structure that cannot be expressed using keywords. We present a novel retrieval model that uses a structured event-based representation. We structure queries and documents as graphs of event mentions and employ graph kernels to measure the query-document similarity. Experimental results on two event-oriented test collections show significant improvements over state-of-the-art keyword-based models.

1 Introduction

The purpose of an information retrieval (IR) system is to retrieve the documents relevant to user's information need expressed in the form of a query. Many information needs are event-oriented, while at the same time there exists an abundance of event-centered texts (e.g., breaking news, police reports) that could satisfy these needs. Furthermore, event-oriented information needs often involve structure that cannot easily be expressed with keyword-based queries (e.g., *"What are the countries that President Bush has visited and in which has his visit triggered protests?"*). Traditional IR models (Salton et al., 1975; Robertson and Jones, 1976; Ponte and Croft, 1998) rely on shallow unstructured representations of documents and queries, making no use of syntactic, semantic, or discourse level information. On the other hand, models utilizing structured event-based representations have not yet proven useful in IR. However, significant advances in event extraction have been achieved in the last decade as the result of standardization efforts (Pustejovsky et al., 2003) and shared evaluation tasks (Verhagen et al., 2010), renewing the interest in structured event-based text representations.

In this paper we present a novel retrieval model that relies on structured event-based representation of text and addresses event-centered queries. We define an *event-oriented query* as a query referring to one or more real-world events, possibly including their participants, the circumstances under which the events occurred, and the temporal relations between the events. We account for such queries by structuring both documents and queries into *event graphs* (Glavaš and Šnajder, 2013b). The event graphs are built from individual event mentions extracted from text, capturing their protagonists, times, locations, and temporal relations. To measure the query-document similarity, we compare the corresponding event graphs using graph kernels (Borgwardt, 2007). Experimental results on two news story collections show significant improvements over state-of-the-art keyword-based models. We also show that our models are especially suitable for retrieval from collections containing topically similar documents.

2 Related Work

Most IR systems are a variant of the vector space model (Salton et al., 1975), probabilistic model (Robertson and Jones, 1976), or language model (Ponte and Croft, 1998), which do not account for associations between query terms. Recent models introduce co-occurrence-based (Park et al., 2011) and syntactic (Shinzato et al., 2012) dependencies. However, these dependencies alone in most cases cannot capture in sufficient detail the semantics of events.

A more comprehensive set of dependencies can be modeled with graph-based representations. Graph-

Proceedings of the TextGraphs-8 Workshop, pages 1–5,
Seattle, Washington, USA, 18 October 2013.

based IR approaches come in two flavors: (1) the entire document collection is represented as a single graph in which queries are inserted as additional vertices (Mihalcea and Tarau, 2004); (2) each query and each document are represented as graphs of concepts, and the relevance of a document for a query is determined by comparing the corresponding graphs (Montes-y Gómez et al., 2000). Our approach fits into the latter group but we represent documents as graphs of events rather than graphs of concepts. In NLP, graph kernels have been used for question type classification (Suzuki, 2005), cross-lingual retrieval (Noh et al., 2009), and recognizing news stories on the same event (Glavaš and Šnajder, 2013b).

Event-based IR is addressed explicitly by Lin et al. (2007), who compare predicate-argument structures extracted from queries to those extracted from documents. However, queries have to be manually decomposed into semantic roles and can contain only a single predicate. Kawahara et al. (2013) propose a similar approach and demonstrate that ranking based on semantic roles outperforms ranking based on syntactic dependencies. Both these approaches target the problem of syntactic alternation but do not consider the queries made of multiple predicates, such as those expressing temporal relations between events.

3 Kernels on Event Graphs

Our approach consists of two steps. First, we construct event graphs from both the document and the query. We then use a graph kernel to measure the query-document similarity and rank the documents.

3.1 Event Graphs

An event graph is a mixed graph in which vertices represent the individual event mentions and edges represent temporal relations between them. More formally, an event graph is a tuple $G = (V, E, A, m, r)$, where V is the set of vertices, E is the set of undirected edges, A is the set of directed edges, $m : V \rightarrow M$ maps the vertices to event mentions, and $r : E \rightarrow R$ assigns temporal relations to edges.

We use a generic representation of a *factual* event mention, which consists of an event anchor and event arguments of four coarse types (*agent*, *target*, *time*, and *location*) (Glavaš and Šnajder, 2013a; Glavaš and Šnajder, 2013b). We adopt the set of temporal relations used in TempEval-2 (Verhagen et al., 2010) (*before*, *after*, and *overlap*), with additional temporal equivalence relation (*equal*).

To build an event graph, we first extract the event mentions and then extract the temporal relations between them. To extract the event anchors, we use a supervised model based on a rich feature set proposed by Glavaš and Šnajder (2013b), performing at 80% F1-score. We then use a robust rule-based approach from Glavaš and Šnajder (2013a) to extract event arguments. Finally, we extract the temporal relations using a supervised model with a rich feature set proposed by Glavaš and Šnajder (2013b). Relation classification performs at 60% F1-score.

To compute the product graph kernels, we must identify event mentions from the query that corefer with mentions from the document. To this end, we employ the model from Glavaš and Šnajder (2013a), which compares the anchors and four types of arguments between a pair of event mentions. The model performs at 67% F-score on the EventCorefBank dataset (Bejan and Harabagiu, 2008).

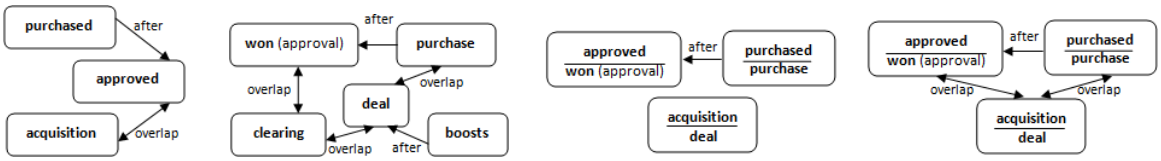
3.2 Product Graph Kernels

Graph kernels provide an expressive measure of similarity between graphs (Borgwardt, 2007). In this work, we use product graph kernel (PGK), a type of random walk graph kernel that counts the common walks between two graphs (Gärtner et al., 2003).

Product graph. The graph product of two labeled graphs, G and G', denoted $G_P = G \times G'$, is a graph with the vertex set

$$V_P = \left\{(v, v') \mid v \in V_G, v' \in V_{G'}, \delta(v, v')\right\}$$

where predicate $\delta(v, v')$ holds iff vertices v and v' are identically labeled (Hammack et al., 2011). Vertices of event graphs have the same label if the event mentions they denote corefer. The edge set of the product is conditioned on the type of the graph product. In the *tensor product*, an edge exists in the product iff the corresponding edges exist in both input graphs and have the same label, i.e., denote the same temporal relation. In the *conormal product*, an edge is introduced iff the corresponding edge exists in at least one input graph. A conormal product may compensate for omitted temporal relations in the input graphs but may introduce spurious edges that do not represent



(a) Query graph (b) Document graph (c) Tensor product (d) Conormal product

Figure 1: Examples of event graphs and their products

true overlap between queries and documents. Fig. 1 shows an example of input graphs and their products.

PGK computation. The PGK for input graphs G and G' is computed as

$$k_{PG}(G, G') = \sum_{i,j=1}^{|V_P|} [(I - \lambda A_P)^{-1}]_{ij}$$

provided $\lambda < 1/d$, where d is the maximum vertex degree in the product graph G_P with the adjacency matrix A_P. In experiments, we set λ to $1/(d+1)$. PGK suffers from tottering (Mahé et al., 2005), a phenomenon due to the repetition of edges in a random walk. A walk that totters between neighboring vertices produces an unrealistically high similarity score. To prevent tottering between neighboring vertices, Mahé et al. (2005) transform the input graphs before computing the kernel score on their product: each edge (v_i, v_j) is converted into a vertex v_e; the edge itself gets replaced with edges (v_e, v_i) and (v_e, v_j). We experiment with Mahé extension for PGK, accounting for the increased probability of one-edge-cycle tottering due the small size of query graphs.

4 Experiments

Test Collections and Queries. To the best of our knowledge, there is no standard test collection available for event-centered IR that we could use to evaluate our models. Thus, we decided to build two such test collections, with 50 queries each: (1) a general collection of topically diverse news stories and (2) a topic-specific collection of news on Syria crisis. The first collection contains 25,948 news stories obtained from EMM News Brief, an online news clustering service.[1] For the topic-specific collection, we selected from the general collection 1387 documents that contain the word *"Syria"* or its derivations.

[1] http://emm.newsbrief.eu

General collection (news stories)
q1: *An ICT giant purchased the phone maker after the government approved the acquisition*
q2: *The warship tried to detain Chinese fishermen but was obstructed by the Chinese vessels*
Topic-specific collection (Syria crisis)
q3: *Syrian forces killed civilians, torched houses, and ransacked stores, overrunning a farmer village*
q4: *Rebels murdered many Syrian soldiers and the government troops blasted the town in central Syria*

Table 1: Example queries from the test collection

For each collection we asked an annotator to compile 50 queries. She was instructed to select at random a document from the collection, read the document carefully, and compile at least one query consisting of at least two event mentions, in such a way that the selected document is relevant for the query. Example queries are shown in Table 1. For instance, query **q1** (whose corresponding event graph is shown in Fig. 1a) was created based on the following document (whose event graph is shown in Fig. 1b):

> *Google Inc. won approval from Chinese regulators for its $12.5 billion purchase of Motorola Mobility Holdings Inc., clearing a final hurdle for a deal that boosts its patents portfolio...*

Relevance judgments. To create relevance judgments, we use the standard IR pooling method with two baseline retrieval models – a TF-IDF weighted vector space model (VSM) and a language model. Our graph-based model was not used for pooling because of time limitations (note that this favors the baseline models because pool-based evaluation is biased against models not contributing to the pool (Büttcher et al., 2007)). Given that EMM News Brief builds clusters of related news and that most EMM

	Model	Collection	
		General	Specific
Baselines	TF-IDF VSM	0.335	0.199
	Hiemstra LM	0.300	0.175
	In_expC2	0.341	0.188
	DFR_BM25	0.332	0.192
Graph-based	Tensor	**0.502**	0.407
	Conormal	0.434	0.359
	Mahé Tensor	0.497	**0.412**
	Mahé Conormal	0.428	0.362

Table 2: Retrieval performance (MAP)

clusters contain less than 50 news stories, we estimate that there are at most 50 relevant documents per query. To get an even better estimate of recall, for each query we pooled the union of top 75 documents retrieved by each of the two baseline models.

One annotator made the relevance judgments for all queries. We asked another annotator to provide judgments for two randomly chosen queries and obtained perfect agreement, which confirmed our intuition that determining relevance for complex event-centered queries is not difficult. The average number of relevant documents per query in the general and topic-specific collection is 12 and 8, respectively.[2]

Results. Table 2 shows the mean average precision (MAP) on both test collections for four graph kernel-based models (tensor/conormal product and with/without Mahé extension). We compare our models to baselines from the three traditional IR paradigms: a TF-IDF-weighted cosine VSM, the language model of Hiemstra (2001), and the best-performing models from the probabilistic Divergence from Randomness (DFR) framework (In_expC2 and DFR_BM25) (Amati, 2003; Ounis et al., 2006). We evaluate these models using the Terrier IR platform.[3]

Overall, all models perform worse on the topic-specific collection, in which all documents are topically related. Our graph kernel models outperform all baseline models ($p<0.01$ for tensor models and $p<0.05$ for conormal models; paired student's t-test) on both collections, with a wider margin on topic-specific than on the general collection. This result

[2]Available at `http://takelab.fer.hr/data`

[3]`http://terrier.org`

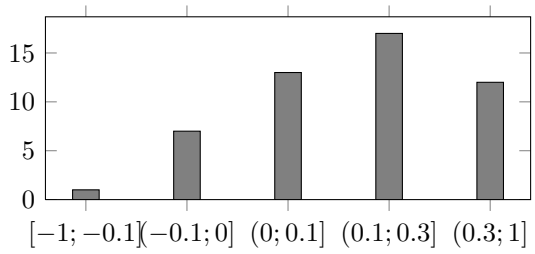


Figure 2: Histogram of AP differences

suggests that the graph-based models are especially suitable for retrieval over topic-specific collections. There is no significant difference between the tensor product and conormal product models, indicating that the conormal product introduces spurious edges more often than it remedies for incorrect extraction of temporal relations. The performance differences due to Mahé extension are not significant, providing no conclusive evidence on the effect of tottering.

To gain more insights into the performance of our event graph-based model, we analyzed per query differences in average precision between our best-performing model (Tensor) and the best-performing baseline (In_expC2) on queries from the general collection. Fig. 2 shows the histogram of differences. Our graph kernel-based model outperforms the baseline on 42 out of 50 queries. A closer inspection of the eight queries on which our model performs worse than the baseline reveals that this is due to (1) an important event mention not being extracted from the query (2 cases) or a (2) failure in coreference resolution between an event mention from the query and a mention from the document (6 cases).

5 Conclusion and Perspectives

We presented a graph-based model for event-centered information retrieval. The model represents queries and documents as event graphs and ranks the documents based on graph kernel similarity. The experiments demonstrate that for event-based queries our graph-based model significantly outperforms state-of-the-art keyword-based retrieval models. Our models are especially suitable for topic-specific collections, on which traditional IR models perform poorly.

An interesting topic for further research is the extension of the model with other types of dependencies between events, such as entailment, causality,

and structural relations. Another direction concerns the effective integration of event graph-based and keyword-based models. We will also consider applications of event graphs on other natural language processing tasks such as text summarization.

Acknowledgments. This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia under the Grant 036-1300646-1986. We thank the reviewers for their comments.

References

Giambattista Amati. 2003. *Probability models for information retrieval based on divergence from randomness*. Ph.D. thesis, University of Glasgow.

Cosmin Adrian Bejan and Sanda Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In *Proc. of the LREC 2008*.

Karsten Michael Borgwardt. 2007. *Graph Kernels*. Ph.D. thesis, Ludwig-Maximilians-Universität München.

Stefan Büttcher, Charles LA Clarke, Peter CK Yeung, and Ian Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. of the ACM SIGIR*, pages 63–70. ACM.

Thomas Gärtner, Peter Flach, and Stefan Wrobel. 2003. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer.

Goran Glavaš and Jan Šnajder. 2013a. Exploring coreference uncertainty of generically extracted event mentions. In *Proc. of the CICLing 2013*, pages 408–422. Springer.

Goran Glavaš and Jan Šnajder. 2013b. Recognizing identical events with graph kernels. In *Proc. of the ACL 2013*, pages 797–803.

Richard Hammack, Wilfried Imrich, and Sandi Klavžar. 2011. *Handbook of Product Graphs*. Discrete Mathematics and Its Applications. CRC Press.

Djoerd Hiemstra. 2001. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata.

Daisuke Kawahara, Keiji Shinzato, Tomohide Shibata, and Sadao Kurohashi. 2013. Precise information retrieval exploiting predicate-argument structures. In *Proc. of the IJCNLP 2013*. In press.

Chia-Hung Lin, Chia-Wei Yen, Jen-Shin Hong, Samuel Cruz-Lara, et al. 2007. Event-based textual document retrieval by using semantic role labeling and coreference resolution. In *IADIS International Conference WWW/Internet 2007*.

Pierre Mahé, Nobuhisa Ueda, Tatsuya Akutsu, Jean-Luc Perret, and Jean-Philippe Vert. 2005. Graph kernels for molecular structure-activity relationship analysis with support vector machines. *Journal of Chemical Information and Modeling*, 45(4):939–951.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proc. of the EMNLP 2004*, volume 4. Barcelona, Spain.

Manuel Montes-y Gómez, Aurelio López-López, and Alexander Gelbukh. 2000. Information retrieval with conceptual graph matching. In *Database and Expert Systems Applications*, pages 312–321. Springer.

Tae-Gil Noh, Seong-Bae Park, Hee-Geun Yoon, Sang-Jo Lee, and Se-Young Park. 2009. An automatic translation of tags for multimedia contents using folksonomy networks. In *Proc. of the ACM SIGIR 2009*, pages 492–499. ACM.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. 2006. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, pages 18–25.

Jae Hyun Park, W Bruce Croft, and David A Smith. 2011. A quasi-synchronous dependence model for information retrieval. In *Proc. of the 20th ACM International Conference on Information and Knowledge Management*, pages 17–26. ACM.

Jay Ponte and Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of the ACM SIGIR*, pages 275–281. ACM.

James Pustejovsky, José Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 3:28–34.

Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2012. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of Information Processing*, 20(1):216–227.

Jun Suzuki. 2005. Kernels for structured data in natural language processing. *Doctor Thesis, Nara Institute of Science and Technology*.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proc. of the SemEval 2010*, pages 57–62.