Queripidia: Query-specific Wikipedia Construction

Laura Dietz
University of Massachusetts
Amherst, MA, U.S.A.
dietz@cs.umass.edu

Michael Schuhmacher and Simone Paolo Ponzetto
University of Mannheim
Mannheim, Germany
{michael,simone}@informatik.uni-mannheim.de

Abstract

With Queripidia we want to answer questions on the fringe of Wikipedia. Given a large, general-purpose knowledge base and a large corpus of web documents, the goal is to consolidate information from both sources to jointly extract material to answer the given query with an automatically compiled query-specific Wikipedia.

1 Introduction

We all refer to Wikipedia when it comes to topics we don't know. But since Wikipedia is created through manual efforts, it is not surprising – but unfortunate – that the scope of its topics and their degrees of details are limited: not every relationship and aspect can be covered and information becomes quickly out-of-date. Consequently, in this work we aim at providing a methodology to respond to users' queries with a set of compiled pages that feel like Wikipedia, in order to enable the users to understand complex topics. For instance, a non profit organizations on fair labor[1] needs to effectively research questions like the following:

- scope and scale of child labor in Turkey;
- Moroccan government's latest efforts to combat forced labor and human trafficking;
- Samsung's history of social compliance or its stance on labor rights.

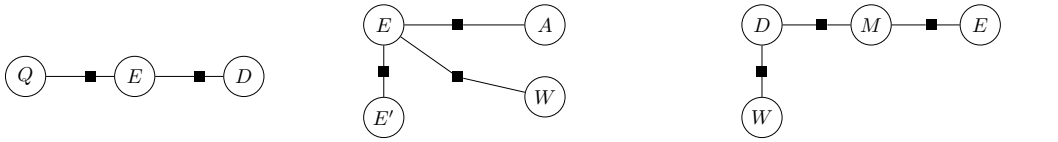
We focus on compiling material from the open Web and structure it with information from knowledge bases like Wikipedia and Freebase. We envision a browsable search interface that informs the user about relevant documents, entities and ontological types, and provides textual evidence for how entities are connected to the query. We aim for serendipitous browsing experiences that refer the user to relevant entities she did not know about and point out connections between entity and query that are surprising yet relevant. We believe that focusing relation extraction to such textual evidence helps to fill in the gaps of the knowledge base in a query-driven manner. That is, we address the problem of incomplete knowledge bases, with a similar motivation to West et al. [9], however we focus primarily on a text-driven approach to answer the user's query.

As a first step towards automatically creating such a query-specific Wikipedia-like site, we need to know (a) what entities to include, and (b) find textual evidences that explain the relationship between the query and the entity – namely, address two tasks that are both related to the text passage retrieval and combination problem introduced by Schlaefer et al. [8]. In this work, we address both these problems by focusing explicitly on finding textual evidence explaining the relationship between query and entity. To solve our task, we make use of a document retrieval system to find relevant text and an entity linking pipeline to establish links between text and entities. Formally, we are solving the following inter-dependent prediction problems.

Problem Statement. Given a query Q,

- rank documents D from the Web collection and text passages thereof by relevance,
- rank entities E from the knowledge base by relevance,
- for each relevant entity, compile a summary explaining its connection to the query.

[1] http://www.verite.org



(a) Joint retrieval problem for given query Q, entity E and document D.

(b) Representing entity E by names A, words W, and neighbors E'.

(c) Representation of documents D, with words W, mentions M and entities E.

Figure 1: Graphical model representation of given query Q, entities E, documents D.

Sauper et al. proposed to compile Wikipedia pages through training templates and their topics [7]. Reasonator[2] renders RDF triples in a human digestible form. In contrast, we focus on identifying textual material for compilation and extraction in open domains. Provenance for relationships between entities is automatically provided through many relation extraction algorithms, e.g., Riedel et al. [6]. In contrast, we focus on provenance for relationships between entities and queries.

We already started some research in the query-specific but open domain setting, exploring the opportunity for joint document and entity retrieval [1]. Focusing on improving document retrieval through latent entities, we achieved significant improvements on well-established retrieval benchmarks Robust04 and ClueWeb12. The approach integrates different indicators from the knowledge base, the entity links in text, and entity-context model [2]. Our previous work focused on document retrieval only. In contrast, this work explores the entity-side of query-specific Wikipedia construction in the following research questions:

Research Question 1. Can we predict a better entity ranking when taking documents and knowledge base into account than through the knowledge base alone?

Research Question 2. Are predicted textual evidences better at explaining the involvement of the entity in the query than the Wikipedia article?

A major success factor in entity-based document retrieval [2] is not only its ability to identify *which* entities are relevant, but also to summarize *how* entities are related to the query. A common misconception is that all aspects of relevant entities are equally important, our approach is to inform this decision through both KB and relevant text.

We include a study on the quality of predicted entities and suggested entity-summaries for immediate presentation to the user. We study both questions through a web prototype called Queripidia which demonstrates the query-specific Wikipedia construction[3]. Our ultimate goal is to submit relevant entity-summaries to a knowledge base construction pipeline to provide more provenance for known relationships and to extract new relationships and new entities in order to populate a knowledge base in a query-driven way.

2 Approach

The overarching idea of this work is to leverage pre-existing knowledge bases, such as from Freebase, to select relevant entities based on the user query. These entities are then used to retrieve relevant documents and text passages from the collection using an entity-inspired retrieval model in combination with the query. We assume that entity mentions in text are already linked to entities in the knowledge base. The retrieved text passages with their entity links are exploited in order to rerank the entities and facts that were originally deemed as being relevant. This joint model between query Q, relevant entity E, relevant document D is depicted in Figure 1a. The factor between entity E and document D is driven through a latent entity representation that governs the entity-inspired retrieval model.

We represent each entity E through different vocabularies: a unique identifier, a distribution over name aliases A, and a distribution over words W. In practice, we combine multiple ways to derive

[2] `http://tools.wmflabs.org/reasonator`

[3] Available online at `http://ciir.cs.umass.edu/~dietz/queripidia/`

such a representation for a given entity. For instance, distributions over words W given E can be derived from the entities' corresponding Wikipedia articles, from context words surrounding entity links with different window sizes and through considering only named entity spans; we leverage approaches from pseudo-relevance feedback to estimate language models for keyword expansion [4]. Likewise, distributions over name aliases A given E can be derived from anchor text of links within Wikipedia, as well as through the mentions of entity links in relevant documents. The combination of different query-independent and query-specific indicators provides a natural smoothing while capturing the entities relations to the query.

The entity representation governs an entity-inspired retrieval model for scoring documents and passages for joint relevance for the query and the entity. The retrieval model probabilistically matches name aliases A and words W to the text part of the document and entity ids to the part with annotated entity links. For probabilistically matching against text, we use the sequential dependence model, a mixture model of unigrams, bigrams and skip-bigrams [5]; the matching model for text and entity links is based on language models with empirical background distributions derived from the corpus.

3 Experimental Evaluation

In previous work [2], our approach lead to significant improved document rankings on standard test data sets Robust04 and ClueWeb12b. In this work we study the model's effectiveness for predicting relevant entities and identifying text passages that explain the entity's involvement in the query. In particular we evaluate different options for ranking entities and predicting entity summaries through the graphical model.

3.1 Setup

Knowledge base. We use a combination of Wikipedia and Freebase as our knowledge base, using dumps from 2012. For each entity in the knowledge base, a document is composed, consisting of a variety of name variants, the full text article, names of in-/out-links. Name variants come from different sources, including the titles of Wikipedia's disambiguation and redirect pages, anchor texts, as well as name aliases from Freebase. These entity-documents are indexed with the search engine Galago[4], retaining entity id information and ontological types.

Corpus. We perform experiments on the ClueWeb12 collection, Category A, which contains 700 million documents. The collection is merged with entity link annotations from the FACC1 dataset [3] and indexed with the search engine Galago.

Queries. We experiment with all 50 queries from the TREC 2013 Web Track, which are gathered from query logs. The queries are general queries such as "dark chocolate health benefits". This test collection comes with a benchmark on relevant documents, but it does not include judgments on entity or passage level.

Annotations. We use our web prototype "Queripidia" to manually create relevance assessments for entities and entity snippets. For every query and entity, we ask a human annotator:

1. Is the entity relevant to answer the query? (RQ1)
2. Do the entity snippets explain how the entity is related to the query? (RQ2)
3. Is the entity's Wikipedia article sufficient to understand how the entity is related to the query? (RQ2 Baseline)

Expecting serendipitous browsing experiences, we ask annotators to re-assess the relevance of the entity after reading the entity snippets. We also collect snippet-level assessments on relevance and spam/junk status. Following the TREC annotation process practice, we assess the top 20 entities (per query) as well as the top 20 passages (per entity) returned by our system. Annotators assess on a binary relevance scale relevant and non-relevant; we also keep track of unjudged entries. To keep the manual assessment effort at an acceptable level, we sampled 17 queries for evaluation; trying to balance simple queries with complex ones and including ambiguous queries such as "rules of golf"

[4] lemurproject.org/galago.php

Method	NDCG@20	num Rel retrieved	P@10
Rank-Fusion	0.748	27.6	0.693
Entity-Context-50	0.694	11.3	0.707
Entity-Context-8	0.690	10.8	0.680
Wikipedia	0.494	18.3	0.500

(a) Entity Relevance

Baseline	Method	top 10	top 20	num Rel retrieved
	Entity-Context-50	0.826	0.806	9.412
	Entity-Context-8	0.801	0.781	8.824
	Rank-Fusion	0.722	0.611	11.471
	Wikipedia	0.478	0.443	4.353
Wikipedia		0.475	0.395	4.529
Rank-Fusion		0.429	0.336	5.882
Entity-Context-8		0.389	0.346	3.471
Entity-Context-50		0.377	0.339	3.588

(b) Entity Snippet Relevance over Wiki article

and "golf gps", we selected query ids 201–206, 208, 210, 220, 223, 224, 228, 234, 236, 237, 239, and 249.

3.2 Result on Entity Ranking (RQ1)

To support Research Question 1, we evaluate different options for ranking entities through the graphical model. We explore the usefulness of a re-estimated entity representation in terms of words W and name aliases A from the contexts surrounding the entities. After a first pass of document retrieval, we extract the entities' contexts, re-estimate the entities' representation, and re-rank the representations (and thereby the entities) by their score under the query. We consider two context windows sizes of 50 and eight words.

We compare this to a baseline that does not consider web documents at all, and instead issues the query to the knowledge base index, and ranks entities accordingly. This is a strong baseline, which led do good performance for the document retrieval problem in our previous work [2]. We also evaluate a combination through an unsupervised rank-based fusion method of all rankings.

We evaluate the rankings with respect to our entity annotations in terms of NDCG@20, Precision@10, and total number of relevant entities retrieved in Table 1a. Our proposed approach achieves much higher P@10 and NDCG@20 scores than the Wikipedia baseline. We find that the Wikipedia baseline and our entity context approach produce complementary rankings. As a result of this, we observe that the simple rank fusion approach is able to achieve the highest NDCG@10. This suggests the potential benefits of applying supervised re-ranking techniques for our task (we leave this for future work).

3.3 Result on Entity Snippets (RQ2)

In Research Question 2, we study methods to extract snippets for relevant entities in order to explain the entity's involvement in the query. We evaluate the added benefit of presenting the snippets to the user. While there exists different snippet re-ranking methods (e.g., based on query terms, lists of entity names, query expansion models), in this study we only evaluate snippet re-ranking through retrieval score under the entity-context expansion model. For different entity ranking methods, we evaluate the fraction of entities in the top 20 (top 10) whose snippet ranking contains relevant snippets (Method). This is compared to the fraction of entities whose Wikipedia article represents a sufficient explanation (Baseline), presented in Table 1b.

The results indicate that on average the entity snippets are always preferred in contrast to the Wikipedia article. Especially for entity rankings derived through the entity context, the snippet representation succeeds in explaining the connection in 80% of the cases. This is in stark contrast to using the Wikipedia article from wiki-rankings, which is successful in less than half of the cases. Inspecting the "Baseline" separately, the results show that when entity rankings are derived from Wikipedia, such articles do indeed provide a good explanation for the entity's involvement. We conclude that rankings and snippets from Wikipedia and the document corpus provide complementary benefits.

3.4 Anecdotal Evaluation

From the query 201 "raspberry pi", which refers to a mini-computer, we learn about his inventor `Eben_Upton`, who developed it in the `United_Kingdom`, while the operating system `Fedora` was customized by software developers in `Toronto, Canada`.

Results for query 206 "wind power" demonstrate how knowledge base and documents are complementary sources of information. In the entity ranking we find Wikipedia articles related to wind power, which mention different countries but not China. These are then used to identify Chinese locations that are related to wind power such as `Qingdao` (wind power location) and `Yinchuan` (manufacturing location). Our approach finally finds textual evidence that `Qingdao` is investing in offshore wind power plants – a fact that is only briefly mentioned in the city's Wikipedia article.

The results for query 222 "male menopause" inform us not only about its medical term `Andropause` and possible hormone therapies, but we also learn its relation to a syndrome named after the novella "Dr. Jekyll and Mr. Hyde", "andropause is a condition where a man goes through an unavoidable change in social interpersonal psychological and even the spiritual aspect similarity of dr jekyll and mr hyde syndrome with male menopause". This query demonstrates some entity linking problems where mentions of "Dr. Jekyll and Mr. Hyde" as the syndrome are linked to the accordingly named novel, movies, and TV series. In the light of this query all these mentions should have been linked to `Dissociative_identity_disorder` instead.

4 Conclusions

In this paper, we presented Queripidia: a first attempt to develop a methodology for automatically compiling Wikipedia-like sites from a back-end knowledge based and document collection based on users' queries. Our approach is capable to jointly reason about compatible query, entity, document, entity summaries, and entity links. We evaluated our method with respect to selecting relevant entities, and providing a summary explaining the entity's involvement. Compared to baselines that only use the knowledge base, we conclude that our joint knowledge base-text approach achieves higher precision and provide better explanations.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by IBM subcontract #4913003298 under DARPA prime contract #HR001-12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] Jeffrey Dalton and Laura Dietz. Constructing query-specific knowledge bases. In *AKBC*, 2013.
[2] Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *SIGIR*, 2014.
[3] Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. FACC1: Freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0), June 2013.
[4] Victor Lavrenko and W. Bruce Croft. Relevance-Based Language Models. In *SIGIR*, 2001.
[5] Donald Metzler and W. Bruce Croft. A Markov random field model for term dependencies. In *SIGIR*, 2005.
[6] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *ECML PKDD*, 2010.
[7] Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *ACL*, 2009.
[8] Nico Schlaefer, Jennifer Chu-Carroll, Eric Nyberg, James Fan, Wlodek Zadrozny, and David Ferrucci. Statistical source expansion for question answering. In *CIKM*, 2011.
[9] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *WWW*, 2014.