Entity-Centric Stream Filtering and Ranking: Filtering and Unfilterable Documents

Gebrekirstos G. Gebremeskel and Arjen P. de Vries

Information Access, CWI, Amsterdam,
Science Park 123, 1098 XG Amsterdam, Netherlands
gebre@cwi.nl
arjen@acm.org

Abstract. Cumulative Citation Recommendation (CCR) is defined as: given a stream of documents on one hand and Knowledge Base (KB) entities on the other, filter, rank and recommend citation-worthy documents. The pipeline encountered in systems that approach this problem involves four stages: filtering, classification, ranking (or scoring), and evaluation. Filtering is only an initial step that reduces the web-scale corpus into a working set of documents more manageable for the subsequent stages. Nevertheless, this step has a large impact on the recall that can be attained maximally. This study analyzes in-depth the main factors that affect recall in the filtering stage. We investigate the impact of choices for corpus cleansing, entity profile construction, entity type, document type, and relevance grade. Because failing on recall in this first step of the pipeline cannot be repaired later on, we identify and characterize the citation-worthy documents that do not pass the filtering stage by examining their contents.

1 Introduction

The maintenance of knowledge bases (KBs) has increasingly become quite a challenge for their curators, considering both the growth of the number of entities considered and the huge amount of online information that appears every day. In this context, researchers have started to create information systems that support the task of Cumulative Citation Recommendation (CCR): given a stream of documents and a set of entities from a Knowledge Base (KB), filter, rank and recommend those documents that curators would consider “citation-worthy”.

KB curators will expect the input stream to cover all the (online) information sources that could contain new information about the entities in the KB, varying from mainstream news sources to forums and blogs. State-of-the-art CCR systems need to operate on web-scale information resources. Current systems therefore divide up their overall approach in multiple stages, e.g., filtering, classification, ranking (or scoring), and evaluation. This paper zooms into this first stage, filtering, an initial step that reduces the web-scale input stream into a working set of documents that is more manageable for the subsequent stages.

Nevertheless, the decisions taken in this stage of the pipeline are critical for recall, and therefore impact the overall performance. The goal of our research is to increase our understanding how design decisions in the filtering stage affect the citation recommendation process.

We build on the resources created in the Knowledge Base Acceleration (KBA) track of the Text REtrieval Conference (TREC), introduced in 2012 with Cumulative Citation Recommendation as the main task. As pointed out in the 2013 track's overview paper [9] and confirmed by our own analysis of participants' reports, the approaches of the thirteen participating teams all suffered from a lack of recall. Could this be an effect of short-comings in the initial filtering stage?

While all TREC-KBA participants applied some form of filtering to produce a smaller working set for their subsequent experiments, the approaches taken vary widely; participants rely on different techniques and resources to represent entities, algorithms may behave differently for the different document types considered in the heterogeneous input stream, and teams use different versions of the corpus. Given these many factors at play, the task of drawing generically applicable conclusions by just comparing overall results of the evaluation campaign seems infeasible. Our paper therefore investigates systematically the impact of choices made in the filtering stage on the overall system performance, varying the methods applied for filtering while fixing the other stages of the pipeline.

The main contributions of the paper are an in-depth analysis of the factors that affect entity-based stream filtering, identifying optimal entity profiles without compromising precision, shedding light on the roles of document types, entity types and relevance grades. We also present a failure analysis, classifying the citation-worthy documents that are not amenable to filtering using the techniques investigated.

The remaining part of the paper is organized as follows. After a brief related work, Section 3 describes the dataset and approach, followed by experiments in Section 4. Sections 5 and 6 discuss their results and a failure analysis. Section 7 summarizes our conclusions.

2 Related Work

Automatic systems to assist KB curators can be seen as a variation of information filtering systems, that "sift through a stream of incoming information to find documents relevant to a set of user needs represented by profiles" [14]. In entity-centric stream filtering, user needs correspond to the KB entities to be curated. However, since the purpose of the filtering component in cumulative citation recommendation is to reduce the web-scale stream into a subset as input for further processing, the decision which documents should be considered citation-worthy is left to later stages in the pipeline.

Other related work addresses the topic of entity-linking, where the goal is to identify entity mentions in online resources and link these to their corresponding KB profiles. Relevant studies include [5, 7], and evaluation resources are developed at the Knowledge Base Population (KBP) track of the Text Analysis

Conference (TAC) [11]. Though related, entity linking emphasizes the problem of locating an entity's mentions in unstructured text, where the primary goal of CCR is to identify an entity's most relevant documents.

Our study is rooted in the research carried out in context of TREC KBA. The problem setup has been essentially the same for both the 2012 and 2013 KBA tracks, but the large size of the 2013 corpus had the effect that all participants resorted to reducing the data-set using an initial filtering stage. Approaches varied significantly in the way they construct entity profiles. Many participants rely on name variants taken from DBpedia, such as labels, names, redirects, birth names, alias, nicknames, same-as and alternative names [15, 6, 12]. Two teams considered (Wikipedia) anchor text and the bold-faced words of the first paragraph of the entity's Wikipedia page [4, 13]. One participant used a Boolean *and* expression built from the tokens of canonical names [8].

Due to the large variety in the methods applied in different stages of the pipeline, it is difficult to infer which approaches are really the best. By focusing on a single component of the pipeline and analyzing the effects of its design choices in detail, we aim at more generally applicable results.

3 Approach

We use the TREC-KBA 2013 dataset[1] to compare the effectiveness of different choices for document and entity representation in the filtering stage. Cleansing refers to pre-processing noisy web text into a canonical "clean" text format. In the specific case of TREC KBA, the organisers provide two versions of the corpus: one that is already cleansed, and one that is the raw data as originally collected by the organisers. Entity profiling refers to creating a representation of the entity based on which the stream of documents is filtered, usually by straightforward matching of their textual contents.

3.1 Dataset Description

The TREC-KBA 2013 dataset consists of three main parts: a time-stamped stream corpus, a set of KB entities to be curated, and a set of relevance judgments. The stream corpus comes in two versions: raw and cleansed. The raw data is a dump of HTML pages. The cleansed version is the raw data after its HTML tags have been stripped off, considering only the documents identified as English (by the Chromium Compact Language Detector[2]). The stream corpus is organized in hourly folders, each of which contains many "chunk files". Each chunk file contains hundreds to hundreds of thousands of semi-structured documents, serialized as thrift objects (one thrift object corresponding to one document). Documents are blog articles, news articles, or social media posts (including tweets). The stream corpus has been derived from three main sources:

[1] `http://trec-kba.org/trec-kba-2013.shtml`

[2] `https://code.google.com/p/chromium-compact-language-detector/`

TREC KBA 2012[3](blogs, news, and urls that were shortened at `bitly.com`), arXiv[4] (e-prints), and spinn3r[5] (blogs).

The KB entities in the dataset consist of 20 Twitter and 121 Wikipedia entities. The entities selected by the organizers of the TREC KBA evaluation are “sparse” (on purpose): they occur in relatively few documents and have an underdeveloped KB entry.

TREC-KBA provides relevance judgments, which are given as document-entity pairs. Documents with citation-worthy content to a given entity are annotated as *vital*, while documents with tangentially relevant content, lacking freshliness or with content that can be useful only for initial KB-dossier creation are annotated as *relevant*. Documents with no relevant content are labeled *neutral*, spam documents are labeled as *garbage*. In total, the set of relevance judgments contains 24162 unique vital-relevant document-entity pairs (9521 vital and 17424 relevant).[6] The relevance judgments have been categorized into 8 source categories: 0.98% arXiv, 0.034% classified, 0.34% forum, 5.65% linking, 11.53% mainstream-news, 18.40% news, 12.93% social and 50.2% weblog. We have regrouped these source categories into three groups, “news”, “social”, and “other”, for two reasons. First, mainstream-news and news are very similar, and can only be distinguished by the underlying data collection process; likewise for weblog and social. Second, some sources contain too few judged document-entity pairs to usefully distinguish between these. The majority of vital or relevant annotations are “social” (63.13%) and “news” (30%). The remaining 7% are grouped as “other”.

3.2 Entity Profiling

The names of the entities that are part of the URL are referred to as their “canonical names”. E.g., entity `http://en.wikipedia.org/wiki/Benjamin_Bronfman` has canonical name “Benjamin Bronfman”, and `https://twitter.com/RonFunchesFor` has canonical name “RonFunchesFor”. For the Wikipedia entities, we derive additional name variants from DBpedia: name, label, birth name, alternative names, redirects, nickname, or alias. For the Twitter entities, we copied the display names manually from their respective Twitter pages. On average, we extract approximately four different name variants for each entity.

For each entity, we create four entity profiles: canonical (cano), canonical partial (cano-part), all name variants combined (all) and their partial names (all-part). Throughout the paper, we refer to the last two profiles as name-variant and name-variant partial, using the terms in parentheses in the Table captions.

[3] `http://trec-kba.org/kba-stream-corpus-2012.shtml`

[4] `http://arxiv.org/`

[5] `http://spinn3r.com/`

[6] The numbers of vital and relevant do not add up to 24162 because some documents are judged as both vital and relevant, by different assessors.

3.3 Evaluation Measures

Our main measure of interest is the recall, as documents missed in this stage cannot be recovered during further processing. We also report the overall performance of a standard high performing setup for the subsequent stages of the pipeline, that we keep constant. Here, we compute the track's standard evaluation metric, max-F, using the scripts provided [9]. Max-F corresponds to the maximally attained F-measure over different cutoffs, averaged over all entities. The default setting takes the vital rating if a document-entity pair has both vital and relevant judgments.

4 Experiments and Results

4.1 Cleansing: Raw or Cleansed

Table 1. Vital recall for cleansed

	cano	cano-part	all	all-part
Wikipedia	61.8	74.8	71.5	77.9
Twitter	1.9	1.9	41.7	80.4
Aggregate	51.0	61.7	66.2	78.4

Table 2. Vital recall for raw

	cano	cano-part	all	all-part
Wikipedia	70.0	86.1	82.4	90.7
Twitter	8.7	8.7	67.9	88.2
Aggregate	59.0	72.2	79.8	90.2

Tables 1 and 2 show that recall (on retrieving each relevance judgment) is higher in the raw version than in the cleansed one. Recall increases on Wikipedia entities vary from 13% to 16.4%, and on Twitter entities from 62.8% to 357.9%. At an aggregate level, recall improvement ranges from 15% to 20.5%. The recall increases are substantial. To put it into perspective, an 15% increase in recall on all entities is a retrieval of 2864 more unique document-entity pairs.

4.2 Entity Profiles

The aggregate recall increase from canonical partial to name-variant partial is 25% and from canonical names to name variants is 35% (see Table 2). This means that a quarter of the documents mentioned the entities by partial names of non-canonical name variants and more than one-third of the documents mention the entities by non-canonical names, respectively. Generally, recall increases as we move from canonicals to canonical partial, to name-variant, and to name-variant partial. The only exception is that using canonical partial leads to a better recall for Wikipedia entities than using the name-variants.

4.3 Relevance Rating: Vital and Relevant

The primary objective of cumulative citation recommendation is to identify the citation-worthy documents. We would like to know if there is a difference between filtering vital and relevant documents (as measured by recall). This could

Table 3. Breakdown of recall performances by document source category

		Aggregate			Wikipedia			Twitter		
		other	news	social	other	news	social	other	news	social
Vital	cano	82.2	65.6	70.9	90.9	80.1	76.8	8.1	6.3	30.5
	cano part	90.4	80.6	83.1	100.0	98.7	90.9	8.1	6.3	30.5
	all	94.8	85.4	83.1	96.4	95.9	85.2	81.1	42.2	68.8
	all part	100	99.2	95.9	100.0	99.2	96.0	100	99.3	94.9
Relevant	cano	84.2	53.4	55.6	88.4	75.6	63.2	10.6	2.2	6.0
	cano part	94.7	68.5	67.8	99.6	97.3	77.3	10.6	2.2	6.0
	all	95.8	90.1	72.9	97.6	95.1	73.1	65.2	78.4	72.0
	all part	98.8	95.5	83.7	99.7	98.0	84.1	83.3	89.7	81.0
All	cano	81.1	56.5	58.2	87.7	76.4	65.7	9.8	3.6	13.5
	cano part	92.0	72.0	70.6	99.6	97.7	80.1	9.8	3.6	13.5
	all	94.8	87.1	75.2	96.8	95.3	75.8	73.5	65.4	71.1
	all part	99.2	96.8	86.6	99.8	98.4	86.8	92.4	92.7	84.9

be helpful to make choices that improve the retrieval of citation-worthy documents selectively. In Table 3, we observe that recall performances considering vital documents only are in general higher than those that consider relevant documents as well. Especially for Wikipedia entities, the vital documents tend to mention the entities by their canonical name. This observation can be explained by the intuition that a highly relevant document usually will mention the entity multiple times, using different forms to refer to it. Those documents are therefore likely to pass the filtering stage.

4.4 Document Categories and Entity Types

The study of recall across document categories (news, social, other) helps us understand how types of documents behave with respect to filtering. Our documents are divided mainly between social and news. Table 3 shows that for Wikipedia entities recall for news documents is higher than for social. In Twitter entities, however, the recall for social documents is higher than for news, except in name-variant partial. Regarding the two types of entities (Wikipedia and Twitter), we see that Wikipedia entities achieve higher recall than Twitter entities (see Tables 1, 2 and 3).

4.5 Impact on Classification

We now will conduct experiments to see how the different choices we made at the filtering stage impact the subsequent steps of the pipeline. Based on the findings of previous work [1, 2, 10], we use a standard pipeline, where the documents passing the filtering stage are classified into their relevance grades. We take the state of the art WEKA's[7] Classification Random Forest and the set of

[7] http://www.cs.waikato.ac.nz/~ml/weka/

features used in [10], for they are small in number, and the resulting classifier is known to be effective for the CCR problem. We follow the official TREC KBA training and testing setting, that is, we train on the number of documents that our filtering system retrieves from the training data and test on those documents retrieved from the test set. For example, when we use cleansed data and canonical profile, we train on training relevance judgments that we retrieve from the cleansed corpus, using the canonical profile, and test on the corresponding test relevance judgments that we retrieve from the cleansed corpus. The same applies for other combinations of choices. In here, we present results showing how the cleansing, entity type, document category, and entity profile impact classification performance.

Table 4. Cleansed: vital max-F

	cano	cano-part	all	all-part
all-entities	0.241	0.261	0.259	0.265
Wikipedia	0.252	0.274	0.265	0.271
twitter	0.105	0.105	0.218	0.228

Table 5. Raw: vital max-F

	cano	cano-part	all	all-part
all-entities	0.240	0.272	0.250	0.251
Wikipedia	0.257	0.257	0.257	0.255
twitter	0.188	0.188	0.208	0.231

Table 6. Cleansed: vital-relevant max-F

	cano	cano-part	all	all-part
all-entities	0.497	0.560	0.579	0.607
Wikipedia	0.546	0.618	0.599	0.617
twitter	0.142	0.142	0.458	0.542

Table 7. Raw: vital-relevant max-F

	cano	cano-part	all	all-part
all-entities	0.509	0.594	0.590	0.612
Wikipedia	0.550	0.617	0.605	0.618
twitter	0.210	0.210	0.499	0.580

Tables 4 and 5 show the max-F performance for vital relevance ranking. On Wikipedia entities, with the exception of canonical entity profiles, the max-F performance using the cleansed version of the corpus is better than that using the raw one. On Twitter entities however, the performance obtained using the raw corpus is better on all entity profiles, with the exception of name-variant partial. This result is interesting, because we saw in previous sections that *recall* when using the raw corpus is substantially higher than using cleansed one. This gain in recall for the raw corpus does however not translate into a gain in max-F for recommending vital documents. In fact, in most cases overall CCR performance decreased. Canonical partial for Wikipedia entities and name-variant partial for Twitter entities achieve the best results. Considering the vital-relevant category (Tables 6 and 7), the results are different. The raw corpus achieves better results in all cases (except in canonical partial of Wikipedia). Summarizing, we find that using the raw corpus has more effect on relevant documents and Twitter entities.

5 Analysis and Discussion

There are 3 interesting observations: 1) cleansing impacts relevant documents and Twitter entities negatively. This is validated by the observation that recall gains in Twitter entities and the relevant categories in the raw corpus also translate into overall performance gains. Cleansing removes more relevant documents than it does vital, which can be explained by the fact that it removes related links and adverts which may contain a mention of the entities. One example we saw was that cleansing removed an image with a text of an entity name which was actually relevant. Cleansing also removes more social documents than news, as can be seen by the fact that most of the missing documents from cleansed are social documents. Twitter entities are affected because of their relation to relevant documents and social documents. Examination of the relevance judgments show that about 70% of relevance judgments for Twitter entities are relevant.

2) Taking both performance (recall at filtering and overall F-score) into account, the trade-off between using a richer entity-profile and retrieval of irrelevant documents results in Wikipedia's canonical partial and Twitter's name variant partial as the two best profiles for Wikipedia and Twitter respectively. This is interesting because TREC KBA participants did not consider Wikipedia's canonical partial as a viable entity profile. Experiments with richer profiles for Wikipedia entities increase recall, but not overall performance.

3) The analysis of entity profiles, relevance ratings, and document categories reveal three differences between Wikipedia and Twitter entities. a) Wikipedia entities achieve higher recall and higher overall performance. b) The best profiles for Wikipedia entities are canonical partial and for Twitter entities name-variant partial. c) The fact that Twitter canonical names achieve very low recall means that documents (specially news and others) almost never use Twitter user names to refer to Twitter entities. However, comparatively speaking, social documents refer to Twitter entities by their user names than news and others suggesting a difference in adherence to standard in names and naming.

The high recall and subsequent higher overall performance of Wikipedia entities can be due to two reasons. First, Wikipedia entities are relatively better described than Twitter entities. The fact that we can retrieve different name variants from DBpedia is an indication of rich description. On the contrary, the fact that the Twitter's richest profile achieves both the highest recall and the highest max-F scores indicates that there is still room for enriching the Twitter entity profiles. Rich description plays a role in both filtering and computation of features such as similarity measures in later stages of the pipeline. By contrast, we have only two names for Twitter entities: their user names and their display names. Second, unfortunately, no standard DBpedia-like resource exists for Twitter entities, from which alternative names can be collected.

In the experimental results, we also observed that recall scores in the vital category are higher than in the relevant category. Based on this result, we can say that the more relevant a document is to an entity, the higher the chance that it will be retrieved with alternative name matching. Across document categories, we observe a pattern in recall of others, followed by news, and then by social.

Social documents are the hardest to retrieve, a consequence of the fact that social documents (tweets and blogs) are more likely to point to a resource where the entity is mentioned, mention the entity with short abbreviation, or talk without mentioning the entities but with some context in mind. By contrast news documents mention the entities they talk about using the common name variants more than social documents do. However, the greater difference in percentage recall between the different entity profiles in the news category indicates news refer to a given entity with different names, rather than by one standard name.

6 Failure Analysis: Vital or Relevant, but Missing

The use of name-variant partial for filtering is an exhaustive attempt to retrieve as many relevant documents as possible, at the cost of bringing in many irrelevant documents. However, we still miss about 2363 (10%) of the vital-relevant documents. If these are not even mentioned by their partial name variants, what type of expressions were they mentioned by?

Table 8 shows the documents that we miss with respect to cleansed and raw corpus. The upper part shows the number of documents missing from cleansed and raw versions of the corpus. The lower part of the table shows the intersections and exclusions in each corpus.

Table 8. The number of documents missing from raw and cleansed extractions (upper part cleansed, lower part raw).

category	Vital	Relevant	Total
Cleansed	1284	1079	2363
Raw	276	4951	5227
missing only from cleansed	1065	2016	3081
missing only from raw	57	160	217
Missing from both	219	1927	2146

One would naturally assume that the set of document-entity pairs retrieved from the cleansed corpus would be a sub-set of those that are retrieved from the raw corpus. We find that this is however not the case; we even find that we retrieve documents from the cleansed corpus that we miss from the raw corpus. Examining the content of the documents reveals that this can be attributed to missing text in the corresponding document representations. Apparently, a (part of) the document content has been lost in the cleansing process, where the removal of HTML tags and non-English content resulted in a loss of partial or entire content. Documents missing from the raw corpus are all social ones (tweets, blogs, posts from other social media), where the conversion to the raw data format (a binary byte array) may have faulted. In both cases, the entity mention happens to be on the part of the text cut out in the transformation.

The most surprising failures correspond to judged documents that do not pass the filtering stage, neither from the raw nor from the cleansed version of the corpus. These may indicate a fundamental shortcoming of filtering the stream using string-matching, requiring potentially more advanced techniques. Our failure analysis identifies 2146 unique document-entity pairs, the majority (86.7%) of which are social documents, 219 of these judged as vital, and related to 35 entities (28 Wikipedia and 7 Twitter).

We observed that among the missing documents, different document ids can have the same content, and be judged multiple times for a given entity.[8] Avoiding duplicates, we randomly selected 35 distinct documents, 13 news and 22 social, one for each entity. Based on this subset of the judgements, we categorized situations under which documents can be vital, without mentioning the entity in ways captured by the entity profiling techniques investigated.

Outgoing link mentions: posts with outgoing links mentioning the entity.

Event place - event: A document that talks about an event is vital to the location entity where it takes place. For example Maha Music Festival takes place in Lewis and Clark_Landing, and a document talking about the festival is vital for the park. There are also cases where an event's address places the event in a park and due to that the document becomes vital to the park. This is basically being mentioned by address which belongs to a larger space.

Entity - related entity: A document about an important figure such as artist, athlete can be vital to another. This is specially true if the two are contending for the same title, one has snatched a title, or award from the other.

Organization - main activity: A document that talks about an area on which the company is active is vital for the organization. For example, Atacocha is a mining company and a news item on mining waste was annotated vital.

Entity - group: If an entity belongs to a certain group (class), a news item about the group can be vital for the individual members. FrankandOak is named innovative company and a news item that talks about the group of innovative companies is relevant for it.

Artist - work: Documents that discuss the work of artists can be relevant to the artists. Such cases include books or films being vital for the book author or the director (actor) of the film. Robocop is film whose screenplay is by Joshua Zetumer. A blog that talks about the film was judged vital for Joshua Zetumer.

Politician - constituency: A major political event in a certain constituency is vital for their politicians. Take e.g. a weblog that talks about two north Dakota counties being drought disasters. The news is considered vital for Joshua Boschee, a politician, a member of North Dakota democratic party.

Head - organization: A document that talks about an entity's organization can be vital: Jasper_Schneider is USDA Rural Development state director for North Dakota and an article about problems of primary health centers in North Dakota is judged vital for him.

[8] For a more detailed analysis of the effect of duplicate documents on evaluation using the KBA stream corpus, refer to [3].

World knowledge, missing content, and disagreement: Some judgments require world knowledge. For example "refreshments, treats, gift shop specials, . . . free and open to the public" is judged relevant to Hjemkomst_Center. Here, the person posting this on social media establishes the relation, not the text itself. Similarly "learn about the gray wolf's hunting and feeding . . . 15 for members, 20 for nonmembers" is judged vital to Red_River_Zoo. For a small remaining number of documents, the authors found no content or could otherwise not reconstruct why the assessors judged them vital.

7 Conclusions

In this paper, we examined the effect of the chain of interactions of cleansing, entity profiles, the effect of the type of entities (Wikipedia or Twitter), categories of documents (news, social, or others) and the relevance ratings (vital or relevant) on recall and overall performance. There is a difference between vital and relevant rankings with respect to filtering: it is easy to achieve higher recall for vital documents only than vital or relevant ones. Given the importance of vital documents (those are the ones we definitely do not want to miss), this is good news for the development of high performing CCR systems.

Cleansing may remove (partial) document content, thereby reducing recall up to 21%. But, this affects the performance of retrieving the relevant documents more than that of vital ones. Looking beyond recall, the overall performance on ranking vital documents improves for Wikipedia entities. Considering also the relevant documents, cleansing affects overall performance negatively. If one is interested in vital documents, then we recommend cleansing, but if one is interested in relevant documents too, then cleansing seems disadvantageous. For KB curation, the emphasis is likely on vital documents, but other tasks (such as filtering information for journalists) may require a high performance on both relevance grades.

Regarding entity profiles, the most effective profiles of Wikipedia entities rely on their canonical partial representation, while the partial name variants perform best for Twitter entities. Because entity type and relevance grade both exhibit differences regarding filtering, they should be dealt with differently to maximize performance. Similarly, social posts and news should be treated differently.

Despite an exhaustive attempt to retrieve as many vital documents as possible, we observe that there are still documents that defy retrieval. About 10% of the vital or relevant documents cannot be identified using our entity profiling techniques, establishing a 90% recall as an upper bound for the full pipeline. The circumstances under which this happens are many. We found that some judged documents are not fully represented in the collection, and in a few cases it is simply not clear why assessors deemed those documents vital. However, the main circumstances under which vital documents can defy filtering can be summarized as outgoing link mentions, venue-event, entity - related entity, organization - main area of operation, entity - group, artist - artist's work, party -

politician, and world knowledge. More advanced entity profiling techniques will be necessary to resolve these situations in the future.

8 Acknowledgments

This study is financed by the COMMIT/ program, as part of the Infiniti project.

References

1. Balog, K., Ramampiaro, H.: Cumulative Citation Recommendation: Classification vs. Ranking. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 941–944 (2013)
2. Balog, K., Ramampiaro, H., Takhirov, N., Nørvåg, K.: Multi-step Classification Approaches to Cumulative Citation Recommendation. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. pp. 121–128 (2013)
3. Baruah, G., Roegiest, A., Smucker, M.D.: The Effect of Expanding Relevance Judgements with Duplicates. In: SIGIR '14 Proceedings of the 37th International ACM SIGIR conference on Research & Development in Information Retrieval. pp. 1159–1162 (2014)
4. Bouvier, V., Bellot, P.: Filtering Entity Centric Documents Using Numerics and Temporals Features within RF Classifier. In: TREC 2013 (2013)
5. Dalton, J., Dietz, L.: A Neighborhood Relevance Model for Entity Linking. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. pp. 149–156 (2013)
6. Dietz, L., Dalton, J.: Umass at TREC 2013 Knowledge Base Acceleration Track. In: TREC 2013 (2013)
7. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285 (2010)
8. Efron, M., Willis, C., Organisciak, P., Balsamo, B., Lucic, A.: The University of Illinois' Graduate School of LIS at TREC 2013. In: TREC 2013 (2013)
9. Frank, J.R., Bauer, J., Kleiman-Weiner, M., Roberts, D.A., Tripuraneni, N., Zhang, C., Ré, C., Voohees, E., Soboroff, I.: Evaluating Stream Filtering for Entity Profile Updates for TREC 2013. In: TREC 2013 (2013)
10. Gebremeskel, G.G., He, J., De Vries, A.P., Lin, J.: Cumulative Citation Recommendation: A Feature-aware Comparisons of Approaches. In: Database and Expert Systems Applications (DEXA). pp. 193–197. IEEE (2014)
11. Ji, H., Grishman, R.: Knowledge Base Bopulation: Successful Approaches and Challenges. In: Proceedings of the 49th Annual Meeting of ACL: Human Language Technologies. pp. 1148–1158 (2011)
12. Liu, X., Fang, H.: A Related Entity Based Approach for Knowledge Base Acceleration. In: TREC 2013 (2013)
13. Nia, M.S., Grant, C., Peng, Y., Wang, D.Z., Petrovic, M.: University of Florida Knowledge Base Acceleration. In: TREC 2013 (2013)
14. Robertson, S.E., Soboroff, I.: The TREC 2002 Filtering Track Report. In: TREC 2012 (2002)
15. Wang, J., Song, D., Lin, C.Y., Liao, L.: BIT and MSRA at TREC KBA Track 2013. In: TREC 2013 (2013)