Building a Domain-Specific Document Collection for Evaluating Metadata Effects on Information Retrieval

Walid Magdy, Jinming Min, Johannes Leveling, Gareth J.F. Jones
Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin, Ireland
{wmagdy, jmin, jleveling, gjones}@computing.dcu.ie

Abstract

This paper describes the development of a structured document collection containing user-generated text and numerical metadata for exploring the exploitation of metadata in information retrieval (IR). The collection consists of more than 61,000 documents extracted from YouTube video pages on basketball in general and NBA (National Basketball Association) in particular, together with a set of 40 topics and their relevance judgements. In addition, a collection of nearly 250,000 user profiles related to the NBA collection is available. Several baseline IR experiments report the effect of using video-associated metadata on retrieval effectiveness. The results surprisingly show that searching the videos titles only performs significantly better than searching additional metadata text fields of the videos such as the tags or the description.

1. Introduction

A huge amount of community-generated data is being created on the web. Web blogs, images, videos, and even Wikipedia are examples of this rapidly growing data. The nature of this data is different in type and content from standard web pages. Finding efficient methods for searching this kind of data becomes a growing demand, and using these resources as language resources has found much attention recently. A video websites such as YouTube[1] is one of these sources of data. The YouTube website contains millions of videos that are uploaded by users to share it with everybody using the web. Videos on YouTube are associated with a huge amount of metadata about the content of the video, and this metadata is generated by the poster of the video (video title and tags). In addition, some different kind of metadata is generated over time about the popularity of the video itself (e.g. rating and number of views).

In this paper, we present design of an information retrieval (IR) test collection based on a snapshot of a collection of video web pages from YouTube. The collection of video pages was crawled in the period before 03/03/2009. The collection is created for the specific domain of basketball to allow research on domain-specific data. All different types of the data and metadata associated with videos were captured and stored.

Several experiments have been conducted to test the usage of different fields of the associated metadata on the effectiveness of IR. The results surprisingly show that using short fields, such as the title of videos, for indexing the data achieves significantly higher retrieval effectiveness than when using additional text that describes the content of the video, such as video tags and description.

The data consists of 61,340 XML documents representing the video pages, 40 topics with corresponding manual relevance assessments to enable IR experiments, and 250k user profiles. These user profiles are for every user who has interacted with the videos (posted the video or commented on it), and it carries some information about users such as, age, gender, location, and may be hobbies. The data structure motivates its usage in various research fields, including information retrieval, adaptive hypermedia, information extraction, ontology generation, sentiment analysis, and others.

The rest of this paper is organized as follows: Section 2 introduces some related work for the usage of this kind of data; Section 3 explains the characteristics of the data and describes the methodology used for crawling it; Section 4 describes the creation of the topics and the relevance assessments; Section 5 discusses the experimental setup and reports the baseline results; Section 6 discusses the applicability of the provided data for video retrieval research; finally, Section 7 concludes the paper and provides the link to obtain the data collection.

2. Related Work

YouTube has always been an interesting source for research about user-generated content in different fields. Some research focused on the efficiency of using this huge amount of data by checking the best methods for caching search results or video links (Cha et al., 2007; Gill et al., 2007). Furthermore, some research has focused on the trend of using the content, such as user behaviour for search and video popularity (Cheng et al., 2008). Other work tested the trend of users' interest by monitoring the

[1] http://www.youtube.com

results of 57 queries on YouTube for eight months about the US 2008 presidential election (Capra et al., 2008). Similar work tried to predict the US presidential election through providing a tool for harvesting videos related to the election, and analysing user interactions with it (Shah and Marchionini, 2007; Shah, 2009).

Different research focused on the effectiveness part of searching YouTube (Yee et al., 2009). The research conducted the current practice to search for videos based on titles, tags, and description. It showed that incorporating the top comment terms into the index processing can improve search accuracy by over 10%. Another work, tried to utilize external resources such as WordNet to enrich the description of annotations of videos in order to get better retrieved results (Altadmiri and Ahmed, 2009). This work showed the effectiveness in correcting some misspelled words in the annotation of videos, which led to improvement in the retrieval recall.

In this paper, we present a new evaluation data collection designed especially for IR, but can be utilized in different research fields as well. The provided collection focuses on one domain of videos only, and is considered much larger than the ones used in (Yee et al., 2009) and (Altadmiri and Ahmed, 2009), which both consist of less than 10,000 documents. The size of data set should lead to better reliable results at least for domain specific research.

3. Data Harvesting

2.1 Data Characteristics

YouTube video pages constitute a very interesting type of data for IR experiments due to their special nature which is characterized by:

1. Short documents (video titles).
2. A range of four types of metadata:
   - Textual (tags, comments, and description).
   - Numerical (video length, average rating; and the number of views, ratings, favorited, comments, and video responses).
   - Relational (video responses, and related videos[2]).
   - Additional (video category, posting user, posting date, and last updating date)

   This allows investigating and evaluating novel IR methods, which make use of metadata.
3. Community-created data that contains ratings, comments, replies, and responses. This enables research in social media in general.

Figure 1 shows the structure of a YouTube video page and all the associated metadata. This combination of varied metadata tags and video content provides a rich environment for multimedia IR research.

[2] Related videos consist of a list of videos which is generated automatically by YouTube. Related videos have content with the current video. Responded videos are videos posted by different users as a response to the current video.

A specific domain was selected for this collection in order to have a focused set of data suitable for research in different fields, which are typically explored separately with different datasets. The basketball NBA domain was chosen in specific because large amounts of suitable content are available on YouTube, including a large number of posted videos and related comments, along with a large community of fans for this type of sport.

2.2 Data Crawling Methodology & Statistics

A total of 61,340 video pages on YouTube were crawled for this collection. Each video page contains the link to the original source page, enabling the possibility of accessing the original video or updating the crawled video page itself. 50 seed queries related to the NBA were used to search the YouTube; these queries were generated from the top 15 ranked NBA players in season 2008, plus the 29 NBA teams along with 6 more NBA general queries such as, e.g. *NBA*, *Michael Jordan*.

The crawling and preparation of the collection was performed following several steps:

1. The top 700 results links for each of the 50 seed queries were extracted. At the time of crawling YouTube search was providing 800 results at most (40 result pages, each contains 20 results), and sometimes the full list are not complete. Therefore, we selected 700 results to assure no failure to our crawler.
2. For each result, the top 20 related video responses and the first 20 related videos links were extracted.
3. All unique video page links were listed.
4. The YouTube page of each video page link was crawled in addition to the gdata pages[3]. Gdata pages are additional pages for the video which are designed for developers, and carry some extra information about the video not appearing on the YouTube page.
5. All metadata fields shown in Figure 1 were extracted, limiting the number of related videos and video responses to 20 each, and the first 500 comments were crawled along the posting user ID
6. For each video page, the extracted data was saved in XML format. Sample is shown in Figure 2.
7. All video pages of other categories other than "Sports" were filtered out
8. The profiles of posting users for all videos and comments were crawled and saved as a separate collection of user profiles consisting of nearly 250,000 unique users.

The period of dates of the posted videos in the collection was found to be between 13/09/2005 and 03/03/2009. Table 1 shows some statistics of the collection including information about the distribution of the values of the metadata associated with the videos.

[3] http://code.google.com/apis/youtube/getting_started.html#data_api

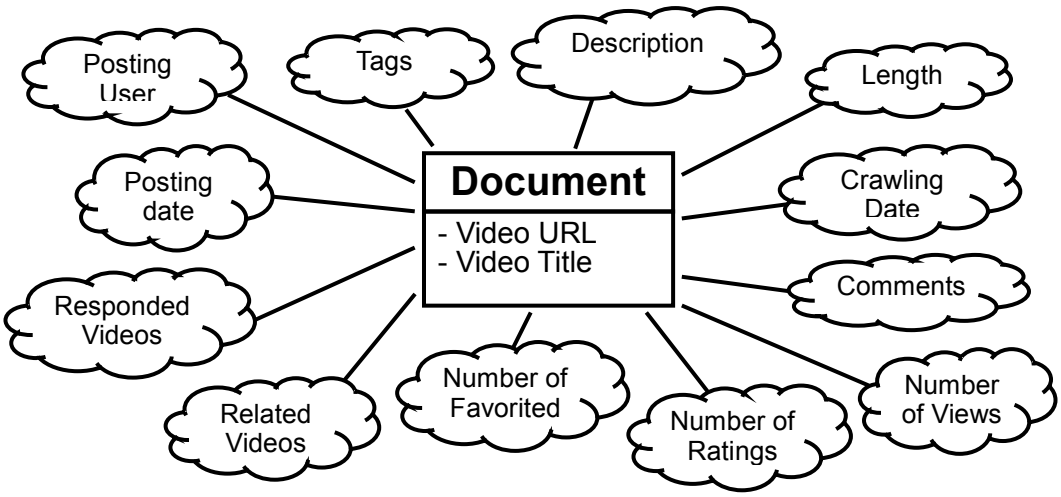

Figure 1: Metadata associated with a YouTube video page

8FHh9Odw9kU.xml

```
<VIDEO link="http://www.youtube.com/watch?v=8FHh9Odw9kU">
    <TITLE> Micheal Jordan Rember the Name</TITLE>
    <LENGTH>229</LENGTH>
    <CATEGORY>Sports</CATEGORY>
    <DESCRIPTION>The Greatest of all time.. </DESCRIPTION>
    <POSTINGDATE>2005-12-28</POSTINGDATE>
    <UPDATED>2009-03-11</UPDATED>
    <POSTUSER>decyrus</POSTUSER>
    <TAGS>
        <tag>Air</tag>
        <tag>Micheal</tag>
        <tag>Jordan</tag>
    </TAGS>
    <VIEWED>145615</VIEWED>
    <RATED avg="4.85">277</RATED>
    <FAVORITED>579</FAVORITED>
    <VIDEORESPONDED>0</VIDEORESPONDED>
    <RELATEDVIDEOS>
        <Video link="http://www.youtube.com/watch?v=MtNEdQ60PMo">
            <title>Michael Jordan Top 10 Moves</title>
            <length>229</length>
            <viewed>17207</viewed>
        </Video>
        <Video link="http://www.youtube.com/watch?v=f6WQLvRvtjs">
            <title>Micheal Jordan vs. Himself</title>
            <length>71</length>
            <viewed>719244</viewed>
        </Video>
        .
        .
    </RELATEDVIDEOS>
    <COMMENTEDON>94</COMMENTEDON>
    <COMMENTS>
        <comment user="dudemanguy">very nice</comment>
        <comment user="AdamHawkey">whats the song called?</comment>
        <comment user="CSFREAK21">THE SONG IS CALLED REMBER THE NAME</comment>
        <comment user="jatt919">jatt</comment>
```

Figure 2: Sample XML document from the YouTube NBA collection

Metadata	Min	Max	Mean	Median	Std Dev
# tags	0	84	12	10	10
# views	0	21,710,757	35,707	3,329	221,091
# comments	0	23,147	58	6	328
Video length	00:00:00	02:38:20	00:02:53	00:02:10	00:02:54
# ratings	0	27,029	52	8	303
Average rating	0	5	4	5	1
# favorited	0	72,230	94	7	687
# video responses	0	232	0	0	2

Table 1: Some statistics of YouTube NBA video pages collection

4. Topics and Relevance Assessment

40 topics on NBA were created to support IR experiments on the NBA collection. These topics were designed to be much more specific than the queries used for crawling the collection. Topic structure follows that of the standard TREC format, with each one consisting of a title, description, and narrative field, in addition to a unique topic ID. A sample topic is shown in Figure 3.

```
<topic>
    <num>4</num>
    <title>Yao Ming missed dunks</title>
    <description>
        Find videos showing one or more missed
        dunks by the Chinese NBA player Yao
        Ming
    </description>
    <narrative>
        A relevant video should contain a missed
        dunk by Yao Ming in any NBA game. The
        video should focus on the missed dunk
    </narrative>
</topic>
```

Figure 3: Sample topic on the YouTube NBA collection

To perform the relevance judgements for each topic, a pooling method was used (Buckley et al., 2006), merging different result lists coming from different search methods. The Lemur search toolkit (Ogilvie and Callan, 2002) was used to search for each query with 5 different models in 4 different indexes. The IR models used to generate the results list were TFIDF, Okapi, Re-rank TFIDF, language modelling (LM) using Dirichlet smoothing, and LM using collection mixture method and Dirichlet smoothing. The detailed description of these retrieval models can be found on the Lemur website [4]. For all IR models, default parameter settings were used.

Four different indexes were built in order to generate varied retrieval results for assessment in the pooling. These included combinations between the textual metadata associated with each video page. Title (TI), tags (TA), description (DE), and related videos titles (RE) were found to be the best textual metadata that describes the content of the video. Indexes were built combining fields as follows: TI, TI+TA, TI+TA+DE, and TI+TA+DE+RE.

The use of the different retrieval models with the four index files led to 20 runs for each query, meaning that 20 different result lists were generated for each topic. The top ranked 60 documents were retrieved for each run. This could potentially result in 1,200 different documents to be judged in the worst case scenario. However, in practice the number of unique documents was far lower, and the union of the 20 results lists led to between 122 and 466 documents to be judged for each topic.

A simple interactive web-based IR application, shown in Figure 4, was developed for relevance assessors to judge the results of each query. The application allows assessors to see the details of each result including video title, tags, description with a link to the video in order to watch before judgment. Other metadata was not shown to the assessor in order to avoid any bias to their decision. The presentation order of retrieval results was randomized in order to ensure a high level of attention by the assessors.

The number of relevant documents (video pages) for the topics was found to range from 1 to 125 documents, with an average of 23 relevant documents per topic.

5. Experimental Setup & Results

Several baseline runs were performed to begin exploration of the effect of searching the collection with different fields in the video page. Topics were used to search four different indexes similar to those created in the pooling step. In order to avoid biased results to any of the models used in pooling, the Indri search toolkit was used to index and search the collection. Indri combines inference network model with language modelling (Metzler and Croft, 2004). Pseudo-relevance feedback (PRF) was tested for its effectiveness on IR, taking 20 as the number of document and feedback terms.

Figures 5 and 6 present the mean average precision (MAP) and recall for the baseline runs for searching the collection with the 4 indexes. The results show that searching only the video titles produces the most precise results, but at the same time the lowest recall. In addition, it was observed that as more text is indexed for each video page, a higher recall is achieved (Figure 6). These results can direct us for using more text in indexes for achieving a higher recall; later re-ranking algorithms could potentially be applied to achieve a higher precision. Both figures illustrate that standard PRF almost always has a negative effect on the precision and recall of the retrieval process for this dataset.

These results can be seen as surprising in the first instance, since they show that using more text in indexes for search leads lower retrieval effectiveness. However, when checking the data, it can be seen that usually the video title contains the most precise short description of the content of the video. The tags and the description fields typically have a very detailed description, but can be noisy because they are user-generated. This can lead to imprecise search results. In addition, using the titles of related videos showed an improvement in the performance when added to the text. However, when combining the video title with the titles of related videos, the result was much lower than using the title alone (0.34 MAP). These results demonstrate that standard IR methods applied to user-generated data might not perform as well as for standard ad-hoc document collections.

[4] http://www.lemurproject.org/

Figure 4: Web tool used for document relevance judgment

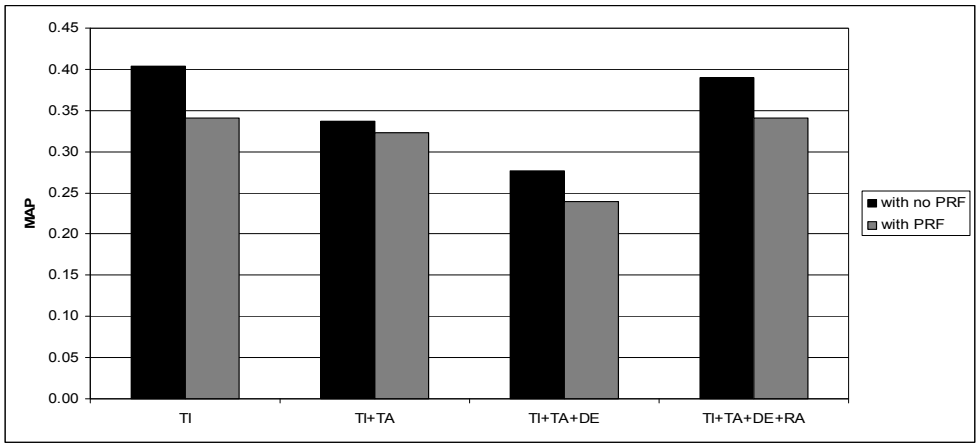

Figure 5: MAP for experiments using different indexed fields of the NBA video page collection with and without pseudo-relevance feedback (PRF)

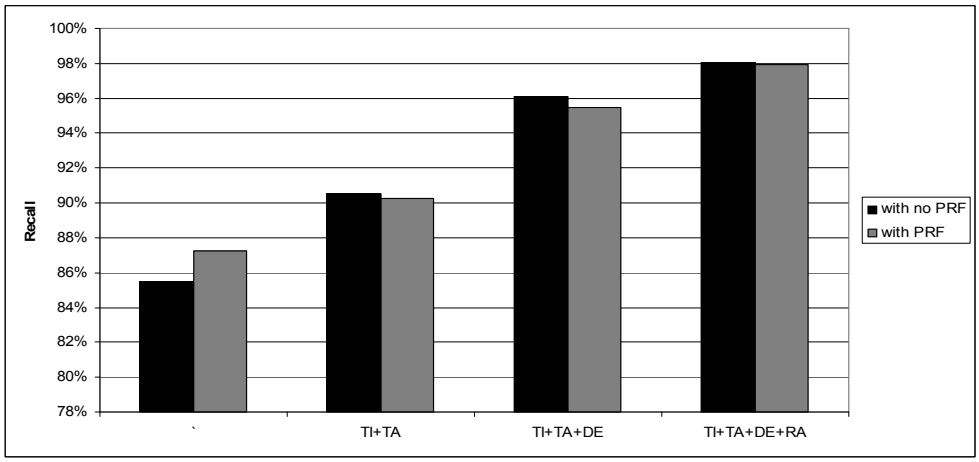

Figure 6: Recall for experiments using different indexed fields of the NBA video page collection with and without pseudo-relevance feedback (PRF)

6. Using the Collection for Video Retrieval

The NBA collection does not contain the videos themselves; it is formed from the video pages that carry all the information on the videos only. Nevertheless, all links of videos are captured for the possibility of using them to download the videos themselves. We performed a quick experiment one year after the crawling of the data to check how many of these video are still available on the web. The experiment showed that 12,714 video pages of the collection have already been removed. The reason of removal from YouTube can be due to terms of use violation, or it can be just removed by the user who posted it. This amount of video pages represents 20% of the data collection.

This experiment shows that using video matching techniques for video retrieval evaluation will be not consistent over time for this type of collections, unless the videos are downloaded at the same time of crawling the collection.

7. Conclusion

In this paper, we described the development of a new benchmark dataset. This has been developed primarily as an IR experimental suite consisting of a document collection, topics, and corresponding relevance assessments. The relevance assessment for this collection is supported by an interactive assessment tool. The data aims at fostering research in diverse areas such as domain-specific IR, video retrieval and text retrieval on structured documents, adaptive hypermedia, and social media search. Baseline results show the effects of using different metadata fields on the retrieval effectiveness.

For future work, the numeric metadata such as video ratings and number of views can be used in conjunction with text search to improve the results. In addition, text fields of the video pages can be used to generate domain-specific named entities. For example, proper nouns such as *Michael Jordan* or *Chicago Bulls* can be easily extracted from the data. This type of named entities can have various usage in different natural language processing (NLP) applications. Finally, an interesting experiment could replicate the full crawling process of the videos in order to examine how the results can change over time.

8. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

We also want to thank all the people involved in creating this new language resource, especially Wei Li and Debasis Ganguly.

9. References

Altadmri A. and A. Ahmed (2009). VisualNet: commonsense knowledgebase for video and image indexing and retrieval application. In: *IEEE International Conference on Intelligent Computing and Intelligent Systems. Shanghai, China*

Buckley C., D. Dimmick, I. Soboroff, and E. Voorhees (2006). Bias and the limits of pooling. In *SIGIR '06, pages 619–620, New York, NY, USA, ACM.*

Capra R. G., C. A. Lee, G. Marchionini, T. Russell, C. Shah, and F. Stutzman (2008). Selection and Context Scoping for Digital Video Collections: An Investigation of YouTube and Blogs. In *JCDL'08, Pittsburgh, Pennsylvania, USA*

Cha M., H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *IMC'07, San Diego, California, USA*

Chenail R. J. (2008). YouTube as a qualitative research asset: Reviewing user generated videos as learning resources. *The Weekly Qualitative Report*, *1*(4), *pages 18-24*.

Cheng, X., C. Dale, and J. Liu (2008). Statistics and Social Network of YouTube Videos. In *Proc. of IWQoS2008, pages 229-238, 2008*

Gill P., M. Arlittz, Z. Li, and A. Mahanti (2007). YouTube Traffic Characterization: A View From the Edge. In *IMC'0, San Diego, California, USA*

Metzler, D. and W. B. Croft (2004). Combining the Language Model and Inference Network Approaches to Retrieval. *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), pages 735-750*.

Ogilvie P. and J. Callan (2002). Experiments using the Lemur toolkit. *TREC 2001. pages. 103-108*.

Shah, C. and G. Marchionini (2007). Preserving 2008 US Presidential Election Videos. In 7th International Web Archiving Workshop, *IWAW'07, Vancouver, British Columbia, Canada*

Shah C (2009). Supporting Research Data Collection from YouTube with TubeKit. In *the Proceedings of YouTube and the 2008 Election Cycle in the United States. Amherst, MA*.

Yee W. G., A. Yates, S. Liu, and O. Frider (2009). Are Web User Comments Useful for Search?. In *Proc. LSDS-IR, 2009, CEUR Workshop Proceedings, Vol. 80. pages. 63-70.*