Information Processing and Management 53 (2017) 799–813

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An in-depth study on diversity evaluation: The importance of intrinsic diversity

Hai-Tao Yu[a,*], Adam Jatowt[b], Roi Blanco[c], Hideo Joho[d], Joemon M. Jose[e]

[a] Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
[b] Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan
[c] IRLab, Computer Science Department, University of A Coruña, Spain
[d] Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
[e] School of Computing Science, University of Glasgow, Glasgow, UK

ARTICLE INFO

Article history:
Received 2 November 2016
Revised 23 February 2017
Accepted 6 March 2017

Keywords:
Extrinsic diversity
Intrinsic diversity
Marginal utility

ABSTRACT

Diversified document ranking has been recognized as an effective strategy to tackle ambiguous and/or underspecified queries. In this paper, we conduct an in-depth study on diversity evaluation that provides insights for assessing the performance of a diversified retrieval system. By casting the widely used diversity metrics (e.g., ERR-IA, α-nDCG and D#-nDCG) into a unified framework based on *marginal utility*, we analyze how these metrics capture *extrinsic diversity* and *intrinsic diversity*. Our analyses show that the prior metrics (ERR-IA, α-nDCG and D#-nDCG) are not able to precisely measure intrinsic diversity if we merely feed a set of subtopics into them in a traditional manner (i.e., without fine-grained relevance knowledge per subtopic). As the redundancy of relevant documents with respect to each specific information need (i.e., subtopic) can not be then detected and solved, the overall diversity evaluation may not be reliable. Furthermore, a series of experiments are conducted on a gold standard collection (English and Chinese) and a set of submitted runs, where the *intent-square metrics* that extend the diversity metrics through incorporating hierarchical subtopics are used as references. The experimental results show that the intent-square metrics disagree with the diversity metrics (ERR-IA and α-nDCG) being used in a traditional way on top-ranked runs, and that the average precision correlation scores between intent-square metrics and the prior diversity metrics (ERR-IA and α-nDCG) are fairly low. These results justify our analyses, and uncover the previously-unknown importance of intrinsic diversity to the overall diversity evaluation.

1. Introduction

Web search engines play an increasingly dominant role in our daily information access. However, generating a high-quality result list in which users can find their desired information from the top few slots is far from being resolved. For example, many users often submit short queries with little or no context, so it is hard to accurately capture their information needs. Thus, merely providing results that satisfy only the most likely information need, will result in dissatisfaction of users with rare information needs. To cope with the ambiguous and/or underspecified queries, the technique of *diversified*

* Corresponding author.
E-mail addresses: yuhaitao@slis.tsukuba.ac.jp (H.-T. Yu), adam@dl.kuis.kyoto-u.ac.jp (A. Jatowt), rblanco@udc.es (R. Blanco), hideo@slis.tsukuba.ac.jp (H. Joho), joemon.jose@glasgow.ac.uk (J.M. Jose).

http://dx.doi.org/10.1016/j.ipm.2017.03.001
0306-4573/

```
▼<topic content="harry potter" id="0083">
 ▼<fls content="harry potter series" poss="0.222222222222">
  ▶<examples>...</examples>
  ▶<sls content="harry potter series_title" poss="0.0762527233115">...</sls>
  ▶<sls content="harry potter series_information" poss="0.0522875816993">...</sls>
  </fls>
 ▼<fls content="harry potter book" poss="0.222222222222">
  ▶<examples>...</examples>
  ▶<sls content="harry potter book_character" poss="0.0936819172113">...</sls>
  ▶<sls content="harry potter book_reading" poss="0.0806100217865">...</sls>
  ▶<sls content="harry potter book_magics" poss="0.0740740740741">...</sls>
  ▶<sls content="harry potter book_scene" poss="0.0566448801743">...</sls>
  ▶<sls content="harry potter book_quotes" poss="0.0305010893246">...</sls>
  </fls>
 ▼<fls content="harry potter film" poss="0.333333333333">
  ▶<examples>...</examples>
  ▶<sls content="harry potter film_watch" poss="0.104575163399">...</sls>
  ▶<sls content="harry potter film_cast" poss="0.0849673202614">...</sls>
  ▶<sls content="harry potter film_information" poss="0.082788671024">...</sls>
  ▶<sls content="harry potter film_music" poss="0.0457516339869">...</sls>
  ▶<sls content="harry potter film_activity" poss="0.0457516339869">...</sls>
  </fls>
 ▼<fls content="harry potter themepark" poss="0.111111111111">
  ▶<examples>...</examples>
  ▶<sls content="harry potter themepark_products" poss="0.0479302832244">...</sls>
  ▶<sls content="harry potter themepark_information" poss="0.0435729847495">...</sls>
  </fls>
 ▼<fls content="harry potter games" poss="0.111111111111">
  ▶<examples>...</examples>
  ▶<sls content="harry potter games_others" poss="0.0348583877996">...</sls>
  ▶<sls content="harry potter games_quiz" poss="0.0261437908497">...</sls>
  ▶<sls content="harry potter games_word game" poss="0.0196078431373">...</sls>
  </fls>
</topic>
```

Fig. 1. Query 0083 *harry potter*.

document ranking has been proposed and has attracted significant attention. In this context, a diversified retrieval system faces a trade-off between relevance and diversity. For a detailed review readers can refer to the works Drosou and Pitoura (2010); Santos, Macdonald, and Ounis (2015) and Vieira et al. (2011).

Effective diversity evaluation provides meaningful insights for assessing a diversified retrieval system, e.g., how well it meets the information needs of users, how to choose among different retrieval models, features, etc. A recent effort for diversity evaluation is the *subtopic based strategy*. The possible information needs underlying a query are represented by a set of subtopics. The number of subtopics that a result list covers and how well a specific subtopic is satisfied provide then the criteria for measuring the *overall diversity* (or *expected diversity*). Note that when using the terms overall diversity and expected diversity, we refer to the diversity that we expect a diversified retrieval system to achieve (sometimes they are used interchangeably). Based on the work by Radlinski, Bennett, Carterette, and Joachims (2009), the overall diversity essentially can be resolved into *extrinsic diversity* and *intrinsic diversity*. Extrinsic diversity corresponds to the problem of addressing the uncertainty about the information needs underlying a query. Intrinsic diversity corresponds to the problem of avoiding excessive redundancy of documents retrieved for a particular information need. The distinction between them is that: Enhancing extrinsic diversity helps to improve the effectiveness of a diversified retrieval system by covering different information needs. The value of enhancing intrinsic diversity helps to increase the satisfaction degree w.r.t. each specific information need. Therefore, extrinsic diversity and intrinsic diversity play different roles in the overall evaluation, both of them are very important.

Take the formal query 0083 *harry potter* from NTCIR-11 (cf. Section 5.1) for example, whose subtopic hierarchy is shown in Fig. 1. The *first-level subtopics*, e.g., *harry potter book* and *harry potter film*, represent different information needs or intents. This clearly reflects the necessity of taking into account the extrinsic diversity, since it is hard to know which first-level subtopic the user is interested in. Given a specific information need, say *harry potter film*, the *second-level subtopics*, e.g., *harry potter film cast* and *harry potter film music* indicate that we should take these different aspects into consideration instead of providing redundant information w.r.t. a single aspect, so as to fully satisfy this information need. This apparently shows the importance of enhancing intrinsic diversity. In the following, the term *subtopic* refers to a first-level subtopic when it is solely used, both of them represent a possible information need or an intent. As a shorthand, we write first-level subtopic and second-level subtopic as *fls* and *sls* respectively.

For effective diversity evaluation, various measures (e.g., *ERR-IA*, α*-nDCG* and *D#-nDCG*) have been proposed. Their properties are further studied and compared by the works Zhai et al. (2003); Clarke et al. (2008); Agrawal, Gollapudi, Halverson,