Identifying Web Pages with Major Contents based on Search Engine Suggests and Topic Modeling

Chen ZHAO†, Yi DING†, Tian NIE††, Jiaqi LI†, Takehito UTSURO†††, Yasuhide KAWADA††, and Noriko KANDO††††

† Graduate School of Systems and Information Engineering, University of Tsukuba 1-1-1, Tennodai, Tsukuba, 305-8573, Japan
†† Logworks Co., Ltd. Tokyo, 151-0051, JAPAN
†††† National Institute of Informatics, Tokyo, 101-8430, JAPAN
††† Graduate School of Systems and Information Engineering, Department of Intelligent Interaction Technologies, University of Tsukuba 1-1-1, Tennodai, Tsukuba, 305-8573, Japan

Abstract This paper addresses the problem of identifying irrelevant items from a small set of similar documents using Web search engine suggests. Specifically, we collected volumes of Web pages through Web search engines and inspected the page contents using topic models. Among each cluster of pages sharing the same topic indicated by the topic model, our technique discovers potential content organization in the current page cluster and identifies pages that are out of focus from that topic. The metrics in our approach mainly consist of search engine suggest frequency and inter-document similarity measures. The intuition is that Web pages collected via the same search queries are more likely to share similar contents. We verify this intuition by implementing a subtopic based document selection framework and making quantitative evaluation against human made labeled data sets. Our evaluation result reveals that suggest frequency analysis along with inter-document similarity measure is effective at filtering off-topic documents in small data sets with satisfactory performance.

Key words search engine suggests, topic model, document filtering, theme aggregation

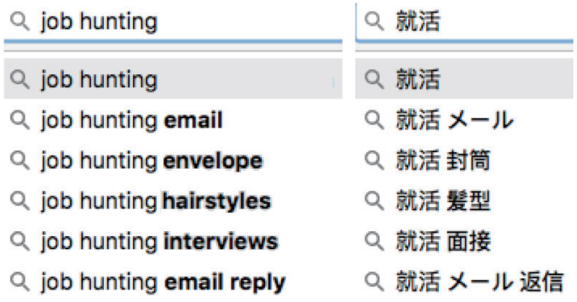

Figure 1 An Example of Search Engine Suggests Provided by the Search Engine

Table 1 Numbers of Collected Suggests and Web Pages for Each Query Focus

Query Focus	就活 (job hunting)	結婚 (marriage)
# suggests	925	947
# web pages	11831	13123

1. Introduction

This paper mainly focuses on analysis techniques that distinguish important documents with major contents from irrelevant ones among a collection of documents all about a specific topic. We first collected search suggests provided by Web search engines about some fixed query focus, which is a predefined keyword on a general topic. Figure 1 shows an example on how a Web search engine provides various search suggests around the query focus keyword "job hunting" ("shyu-katsu" in the Japanese case). These suggests serve as indication of frequent user search logs and trending topics over the time. We used all the collected search suggests as queries to extract Web pages from the search engine meanwhile keeping track of which queries were used to acquire every page and saving them as page-wise information for later stage analysis. For example, in the case we query the search engine with the term "job hunting hairstyles" ("shyu-katsu kamigata" for Japanese), we attach every Web page collected from this query with the phrase "job hunting

hairstyles" as one of its search engine suggests. Next we apply all the Web pages regarding a query focus which is "job hunting" for this example to the LDA topic model. The LDA topic model is a popular topic model among text processing community and it performs statistical sampling by Latent Dirichlet Allocation [3] to model topic distributions to a collection of documents. Based on the topic distribution from the LDA model, we assign every collect Web page a topic of maximum probability so that our Web page collection is rearranged into a fixed number of topic clusters, which will be explained in details in later sections. Our primary objective is to design on top of the LDA topic assignment a framework that automatically analyzes Web pages in each topic cluster, so that pages with major subtopics will be identified versus pages with minor subtopics. In this paper a subtopic within an LDA topic cluster will always be referred to as a "theme" to avoid confusion. We define major themes as page contents shared by at least some number of pages in the same topic cluster. The reason of such a definition stems from the sense that on-focus themes tend be more frequently covered by Web pages while themes with rare occurrence are more likely to be trivial contents. The purpose of this framework is to identify on-focus pages in a topic cluster and filter out less relevant trivial contents — pages with minor themes that are not semantically close enough to all the other pages. This paper presents all its analysis on two different query focuses that cover various Japanese query results about job hunting and marriage issues. The number of total suggests collected and Web pages are listed in Table 1. The rest of the paper is organized as follows. First it introduces some details on search engine suggest assignment and LDA topic model structure. Then our main approach is explained along with the criteria how human made reference data is generated, followed by the evaluation results of the approach. Next is a brief glimpse of previous related research. Finally, it comes to conclusion.

2. Collecting Search Engine Suggests and Web Pages

2.1 Collecting Search Engine Suggests

For a given query focus keyword, we specify about 100 types of Japanese hiragana characters to Google® search engine from which we then collect not exceeding 1000 suggests. These 100 types of Japanese hiragana characters include Japanese alphabet consisting of 50 characters, voiced and semi-voiced variants of voiceless characters and Youon (a variation of diphthong as language feature in Japanese). For example, once we type "shu-katsu a" ("job hunting a") into the search field, a list of suggests are popped out all starting with the reading character "a" such as "aisatsu" ("greeting") and "anata no tsuyomi" ("your strengths").

2.2 Collecting Web Pages

We used Yahoo! Search BOSS API(注1) to scrape web pages from the search engine. Using the web search engine suggests we collected in the previous section combined with the query focus keyword as queries (in the form of AND search), we always collect the first 20 pages returned per query. The set of web pages queried by suggest s can be represented as $D(s, N)$ where N is 20 as a constant standing for the top N pages. As previously mentioned we save the search engine suggests for every Web page. Since different search engine suggests could lead to the same Web page, one single Web page could have multiple suggests. So we maintain a suggest set $\mathbb{S}(d)$ for each Web page d, so that $\mathbb{S}(d)$ contains all the suggests that were used to search the page d. Therefore suggests of a web page are saved as follows.

$$\mathbb{S}(d) \;=\; \Big\{ s \in S \Big| d \in D(s, N) \Big\}$$

3. LDA Topic Model

3.1 Topic Model

This paper employs LDA (Latent Dirichlet Allocation) [3] to model topic distributions among documents. Given a preset constant K representing the number of output topics, The LDA topic model takes a collection of documents (in our case Web pages for one query focus), treats every single document as a sequence of words and estimates the word distribution $p(w \,|\, z_n)$ $(w \in V)$ for every topic z_n $(n = 1, \ldots, K)$ as well as a topic distribution $p(z_n \,|\, d)$ for every document where V is the vocabulary set(注2). This paper adopts GibbsLDA++(注3) as the toolkit while the parameters are tuned through a preliminary

evaluation by examining the number of topics as $K = 40$ and 50 for query focuses "marriage" and "job hunting", respectively.

3.2 Assigning a Topic to a Web Page

Let D be the document set containing all collected Web pages and K be the number of topics. Given the topic model is applied, we have a topic distribution $p(z_n \,|\, d)$ available for every $d\,(d \in D)$. We then assign every document d the topic with the highest probability among all its $p(z_n \mid d)$. The following formula defines this process.

$$D(z_n) \;=\; \Big\{ d \in D \;\Big|\; z_n = \underset{z_u \;(u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \Big\}$$

(注1) : http://developer.yahoo.com/search/boss

(注2) : In this paper, as the set V of vocabulary, we use the set of entry titles of the Japanese version of Wikipedia, where the version we used in this evaluation was downloaded in March 2014 and has about 1,407,000 entries.

(注3) : `http://gibbslda.sourceforge.net/`

Table 2 Numbers of Suggests satisfying Lower Bounds of Frequency

Suggest Frequency Lower Bound	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Syu-katsu (Job Hunting)	921	846	692	592	495	425	379	322	264	224	187	132	88	53	30	12	3
Kekkon (Marriage)	946	938	884	798	706	604	533	453	376	305	242	182	132	88	50	27	8

The net effect is that for every topic z_n, there are a group of corresponding documents that are assigned to z_n. Since we never assign multiple topics to the same document, there is no overlap between topic clusters $D(z_n)$ for $n = 1, \ldots, K$.

4. Identifying Documents of Major Contents

4.1 Overview

Our framework is designed in such a way that within each topic cluster $D(z_n)$ it selects documents of on-focus themes and leaves the rest as irrelevant theme documents. We name the set of such selected documents as "major contents" for topic cluster $D(z_n)$. We define a document $d\,(d \in D)$ to be a major content document given that in the current topic cluster $D(z_n)$, there are at least d_f documents about the same theme as d, including d itself. In other words, document d is said to belong to major contents if we are able to find $d_f - 1$ other documents sharing one or more themes within $D(z_n)$. The intuition of this definition comes from the assumption that if a majority of documents covers some theme then this theme is likely to be on-focus content otherwise it tends to be trivial. This section first introduces how manually certified reference data set is produced. Next it explains how suggest frequency helps achieve the above semantics. It then discusses similarity measures used to help recheck and improve the document selection results.

4.2 Creating Reference Data

The reference data set is a sequence of manual labels independently created by purely manual work. This data set serves as a standard solution to the document selection problem for evaluation purpose. First, considering the practical manual workload, we restrain the reference data set within the top r documents in each topic cluster ranked by topic model probability $p(z_n\,|\,d)$.

$$D_{rank}(z_n, r) = \Big\{ d \in D(z_n) \;\Big|\; p(z_n\,|\,d) \text{ ranks at least } r \text{ in } D(z_n) \text{ by descending order} \Big\}$$

Then, complying with the definition of "major contents" specified in section 4.1, human readers go through every $D(z_n)$ determining the themes contained in every document based on real-life semantics and count the number of documents containing each of the themes. If at least two other documents in the same cluster are found to contain similar information with d by common sense, d is labeled to be in major contents. Thus, the actual reference data selects documents that belong to major contents from the top r documents as follows.

$$D_{ref}(z_n, r, d_f) = \Big\{ d \in D_{rank}(z_n, r) \;\Big|\; \text{at least } d_f \text{ documents in } D_{rank}(z_n, r) \text{ contain the same theme with } d \Big\}$$

4.3 In-topic Suggest Frequency Based Document Selection Method

The first part of our approach relies on suggest frequency count to determine whether a document belongs to major contents with respect to its topic cluster. This method attempts to reproduce the manual work in section 4.2 in an automatic way by assuming that a document queried by some suggest always covers contents relevant to that suggest, which is true only for non-spam Web pages. Furthermore, search engine suggests with higher counts of occurrence are supposed to be more reliable in terms of revealing document themes because a sufficient number of documents are available to help verify this assumption.

As previously stated, every document is coupled with a list of Web search engine suggests as a record about which queries that page was retrieved with. Therefore within every topic cluster $D(z_n)$ we count the occurrences of every suggest that belongs to documents in $D(z_n)$. Table 2 lists overall statistics on total number of search engine suggests with frequency above different thresholds within topics. Formally this frequency is defined by the following formula.

$$f(s, z_n) = \Big| \Big\{ d \in D(z_n) \mid s \in S(d) \Big\} \Big|$$

Here $S(d)$ is the set of all search engine suggests for document d. An alternative explanation of suggest frequency $f(s, z_n)$ is the number of documents in $D(z_n)$ containing suggest s. Furthermore, we define the maximum suggest frequency of d, $f_{max}(d)$ as the most frequent suggest among $S(d)$.

$$f_{max}(d) = \Big| \operatorname*{argmax}_{s \in S(d)} f(s, z_n) \Big|$$

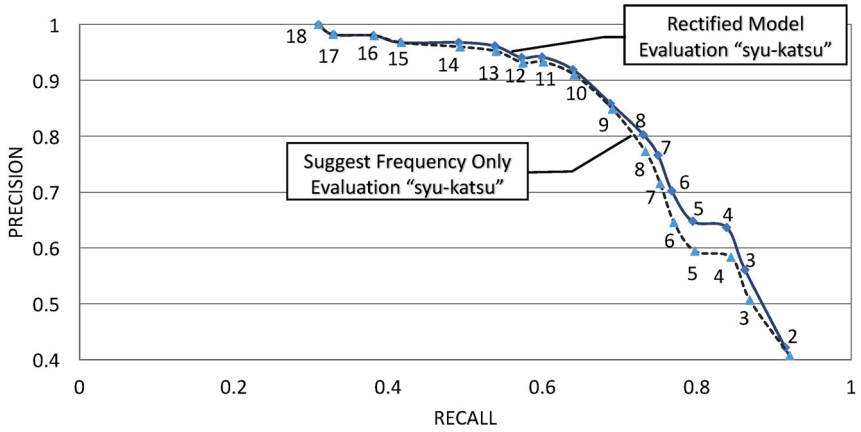

Figure 2 Query Focus Syu-katsu (Job Hunting) Macro Precisions and Recalls by varying

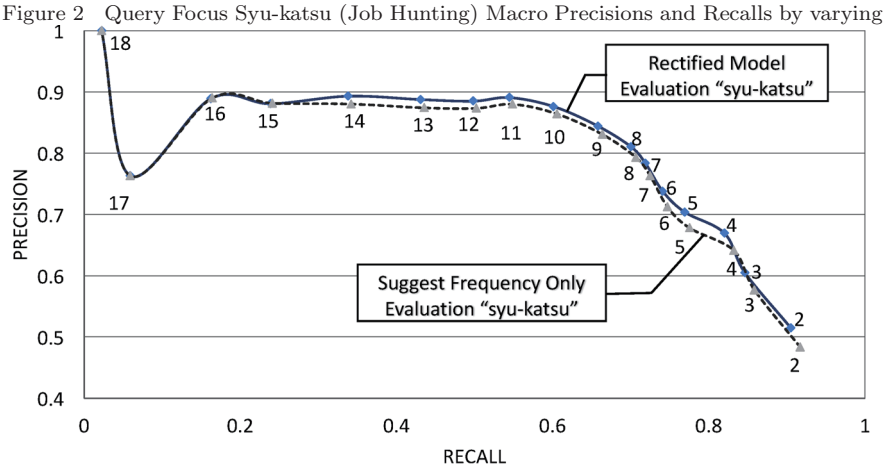

Figure 3 Query Focus Syu-katsu (Job Hunting) Micro Precisions and Recalls by varying Suggest Frequencies

By the above definition, we are able to compute the maximum suggest frequency of every document in separate clusters $D(z_n)$. Our expectation is that document suggests $S(d)$ would well indicate document themes. For example, if $D(z_n)$ contains "hairstyle" and "email reply", we infer that document d covers themes on both hairstyle tips and email communication issues in its context. Whereas the maximum suggest $f_{max}(d)$ reflects the maximum number of possible documents in $D(z_n)$ sharing identical suggest(s) with d, we approximate the semantics on major contents by selecting documents whose $f_{max}(d)$ satisfies some lower bound f_{lbd} as defined below.

$$D_f(z_n, f_{lbd}) = \left\{ d \in D(z_n) \mid f_{max}(d) \geqq f_{lbd} \right\}$$

Then, we restrict the evaluation data within the top r documents in each topic cluster ranked by topic model probability $p(z_n \mid d)$ as the set $D_f(z_n, f_{lbd}, r)$ given below:

$$D_f(z_n, f_{lbd}, r) \;=\; D_f(z_n, f_{lbd}) \cap D_{rank}(z_n, r)$$

For every topic cluster $D(z_n)$, we select $D_f(z_n, f_{lbd}, r)$ to be major content documents and all the rest to be trivial ones that are filtered from major contents as relatively irrelevant documents. In addition, we define the set of topics with non-empty major contents as $T_f(f_{lbd})$, so that topics with empty major contents will be treated as pure garbage clusters.

$$T_f(f_{lbd}) = \left\{ z_n \middle| D_f(z_n, f_{lbd}, r) \neq \emptyset \right\}$$

4.4 Incorporating In-topic Document Similarity Measures

Although the assumption in the previous section that suggests serve as hints about themes in a document is true for most Web pages in our dataset, there exist a lot of spam and erroneous pages where this assumption does not stand as real world Web text is mostly messy. Therefore we need to further examine documents by distance measures to verify whether a document is truly sharing common characteristics with the others to be eligible for major contents. This paper utilizes weighted Jaccard distance [6] for computing correlation between documents. The distance between document d and d' is defined as follows.

$$Jaccard(d, d') \;=\; 1 \;-\; \frac{\sum_{w \in d \bigcup d'} \min(d_w, d'_w)}{\sum_{w \in d \bigcup d'} \max(d_w, d'_w)}$$

where d_w and d'_k are count of word w in document d and

$d^{'}$. Then we take the similarity between d and d' is computed as $1 - Jaccard(d, d')$ as

$$Sim(d, d') \ = \ \frac{\sum_{w \in d \bigcup d'} \ \min(d_w, d'_w)}{\sum_{w \in d \bigcup d'} \ \max(d_w, d'_w)}$$

Based on the above measures, we select d from $D(z_n)$ as a major content document if d satisfies the condition that there are at least $d_f - 1$ other documents in $D(z_n)$ above some similarity threshold s_{lbd}. Furthermore, all inter-document similarities are computed within the top r documents $D_{rank}(z_n, r)$ to simulate the reference data set. We define all such documents similar to d as the nearest neighbor set N expressed as

$$N(d, z_n, s_{lbd}) \ = \ \left\{ d' \in D_{rank}(z_n, r) \ \middle| \ d' \neq d, \ Sim(d, d') \geqq s_{lbd} \right\}$$

The major contents $D(d, z_n, d_f, s_{lbd})$ is then selected by D_s based on the nearest neighbor set size.

$$D_s(z_n, s_{lbd}, d_f, r) \ = \\ \left\{ d' \in D_{rank}(z_n, r) \ \middle| \ |N(d, z_n, s_{lbd}, r)| \geqq d_f - 1 \right\}$$

The above similarity measure scheme can be stated in plain words that only if there exist a least number of neighbor documents close enough to the current one, can the current document be recognized as part of major contents.

Now that we have defined how similarity measures certify major contents as $D_s(z_n, s_{lbd}, d_f)$, we now combine this technique with the frequency count method in the previous section to generate the final output. We apply our model upon each topic $D(z_n)$. The model selects documents satisfying both suggest frequency and similarity criteria as the set $D(z_n, f_{lbd}, d_f, s_{lbd})$, which is the intersection of the documents by both methods. For every LDA topic cluster z_n,

$$D(z_n, f_{lbd}, d_f, s_{lbd}, r) = D_s(z_n, s_{lbd}, d_f, r) \bigcap D_f(z_n, f_{lbd}, r)$$

is the final output from the model and will get evaluated against the reference data set $D_{ref}(z_n, r, d_f)$.

Similar to $T_f(f_{lbd})$ in the previous section, the set of topics with non-empty major contents in the final model is defined as $T(f_{lbd})$

$$T(f_{lbd}) = \left\{ z_n \middle| D(z_n, f_{lbd}, d_f, s_{lbd}, r) \neq \emptyset \right\}$$

5. Evaluation

5.1 Procedures

This paper experiments the proposed model on Web page collections of two query focuses as listed in Table 1. The topic model has a preset output topic number $K = 40$ for "job hunting" and $K = 50$ for "marriage". For every topic cluster $D(z_n)$, the model output $D(z_n, f_{lbd}, d_f, s_{lbd}, r)$ is evaluated against the labeled reference $D_{ref}(z_n, r, d_f)$(注4) so that we compute the precision and recall for each z_n based on whether a document is correctly classified as part of major contents. With evaluation outcomes from each individual topic cluster, we integrate them to produce the final macro/micro precisions and recalls for the entire query focus. The final macro/micro precision and recall for a query focus are defined as follows.

$$Macro\ Recall(f_{lbd}) \ = \\ \frac{\sum_{z_n \in T(f_{lbd})} \frac{\left| D_{ref}(z_n, r, d_f) \bigcap D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}{\left| D_{ref}(z_n, r, d_f) \right|}}{\left| \left\{ z_n (n = 1, \ldots, K) \ \middle| \ D_{ref}(z_n, r, d_f) \neq \emptyset \right\} \right|}$$

$$Macro\ Precision(f_{lbd}) \ = \\ \frac{\sum_{z_n \in T(f_{lbd})} \frac{\left| D_{ref}(z_n, r, d_f) \bigcap D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}{\left| D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}}{\left| T(f_{lbd}) \right|}$$

$$Micro\ Recall(f_{lbd}) \ = \\ \frac{\sum_{z_n \in T(f_{lbd})} \left| D_{ref}(z_n, r, d_f) \bigcap D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}{\sum_{z_n (n=1,\ldots,K)} \left| D_{ref}(z_n, r, d_f) \right|}$$

$$Micro\ Precision(f_{lbd}) \ = \\ \frac{\sum_{z_n \in T(f_{lbd})} \left| D_{ref}(z_n, r, d_f) \bigcap D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}{\sum_{z_n \in T(f_{lbd})} \left| D(z_n, f_{lbd}, d_f, s_{lbd}, r) \right|}$$

5.2 Evaluation Report

The final document selection model accepts 5 parameters to generate major contents $D(z_n, f_{lbd}, d_f, s_{lbd}, r)$ as the model output. r is set to be 30. As noted before, d_f is constantly 3. s_{lbd} is pre-modulated to 0.10 for query focus "job hunting" and 0.15 for "marriage". With d_f and s_{lbd} fixed, the model is applied for multiple runs with different suggest frequency thresholds for easier observation on how the model performance varies with different f_{lbd}. Figure 2 to Figure 5 showcase precisions and recalls corresponding to f_{lbd} ranging from 2 to 18 for both query focuses. In addition to final model output $D(z_n, f_{lbd}, d_f, s_{lbd}, r)$, Figure 2 to Figure 5 also include evaluation on suggest frequency based

(注4) : In this paper, d_f is constant and $d_f = 3$ for all cases.

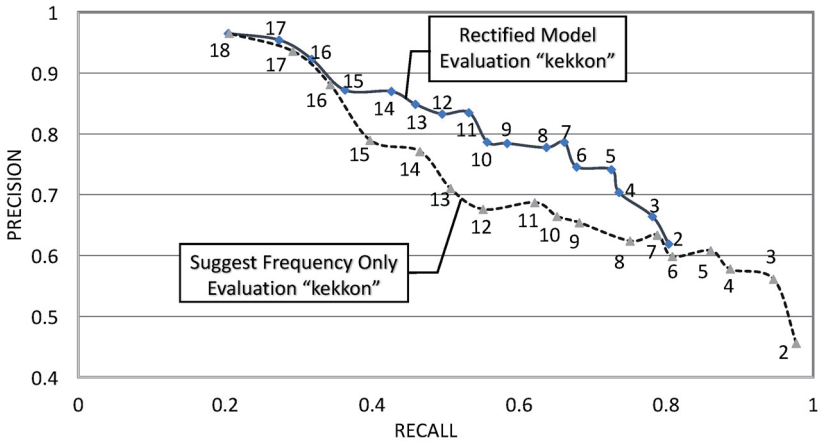

Figure 4 Query Focus Kekkon (Marriage) Macro Precisions and Recalls by varying Suggest Frequencies

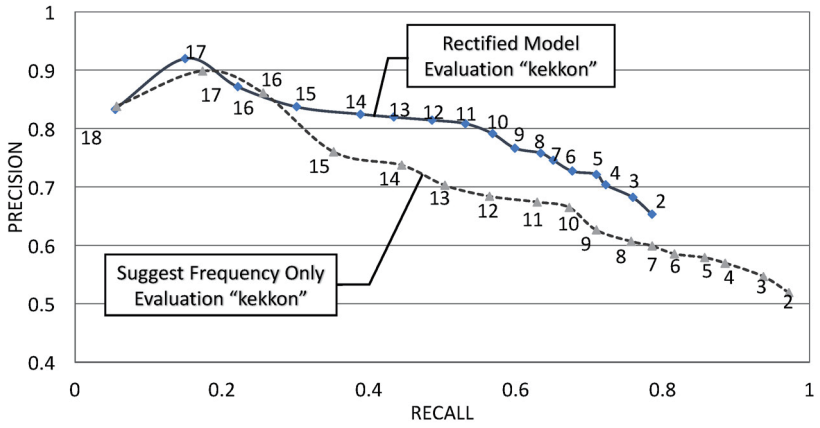

Figure 5 Query Focus Kekkon (Marriage) Micro Precisions and Recalls by varying Suggest Frequencies

selector $D_f(z_n, f_{lbd}, r)$ without rectification by any similarity measures, in order to confirm how much improvement similarity measures contribute to. The precision and recall measures of $D_f(z_n, f_{lbd}, r)$ are defined in the same approach as final model output. $D_f(z_n, f_{lbd}, r)$ evaluations are presented in dashed lines figures and defined as Macro/Micro $Precision_f/Recall_f$ below.

$$Macro\ Recall_f(f_{lbd}) = \frac{\sum_{z_n \in T_f(f_{lbd})} \frac{\left|D_{ref}(z_n, r, d_f) \bigcap D_f(z_n, f_{lbd}, r)\right|}{\left|D_{ref}(z_n, r, d_f)\right|}}{\left|\left\{z_n(n = 1, \ldots, K) \mid D_{ref}(z_n, r, d_f) \neq \emptyset\right\}\right|}$$

$$Macro\ Precision_f(f_{lbd}) = \frac{\sum_{z_n \in T_f(f_{lbd})} \frac{\left|D_{ref}(z_n, r, d_f) \bigcap D_f(z_n, f_{lbd}, r)\right|}{\left|D_f(z_n, f_{lbd}, r)\right|}}{\left|T_f(f_{lbd})\right|}$$

$$Micro\ Recall_f(f_{lbd}) = \frac{\sum_{z_n \in T_f(f_{lbd})} \left|D_{ref}(z_n, r, d_f) \bigcap D_f(z_n, f_{lbd}, r)\right|}{\sum_{z_n(n=1,\ldots,K)} \left|D_{ref}(z_n, r, d_f)\right|}$$

$$Micro\ Precision_f(f_{lbd}) = \frac{\sum_{z_n \in T_f(f_{lbd})} \left|D_{ref}(z_n, r, d_f) \bigcap D(z_n, f_{lbd}, r)\right|}{\sum_{z_n \in T_f(f_{lbd})} \left|D_f(z_n, f_{lbd}, r)\right|}$$

5. 3 Examples of Major Content Selection

This section describes a concrete use case example on how the document selection scheme practically works. Details are depicted in Figure 6. This example contains a topic cluster $D(z_n)$ for query focus "marriage". This topic contains documents mainly focusing on common issues to consider when choosing marriage partners such as occupations, incomes and personality. The right column lists the model output. Since

documents with ID 1, 2, 3, 4 share the same theme about influence of occupations on marriage they are selected as major contents. Similar case goes for ID 5, 6, 7 which all cover contents on qualifications for marriage, especially for men. Remaining documents such as 8, 9, 10 are not major contents either because that they share no common theme with at least 2 others in the cluster or that their suggest frequencies are below the preset threshold $f_{lbd} = 3$. For this topic alone, the recall is calculated as the proportion of correctly selected documents in the reference data (the reference set is not shown). The precision in this case is the proportion of documents in the reference set among those selected by the proposed method.

6. Related Work

The major content selection scheme presented in this paper finds its insight mostly from existing topic modeling metrics. Blei et al. [3] proposes perplexity as a topic quality metric. de Wall and Barnard [7] discusses the problem of vocabulary dependence in perplexity and raise the concept "topic stability" as an evaluation technique. Similarly there are other works [9], [15] studying correctness of topic modeling. This paper differs from those works by focusing on in-topic contents instead of overall probabilistic behavior of the topic model.

As for in-topic content evaluation, there also exist quite a few quality evaluation schemes that measure how meaningful topics are including Chang et al. [5] using manual evaluation on held-out keywords and other work [1], [4], [10]?[12], [14] that rely on external knowledge resources such as Wikipedia, WordNet(注5), search engine information and news corpora. Lau et al. [8] and Röder et al. [13] do some systematic comparison among various topic evaluation techniques. AlSumait et al. [2] recognizes junk topics through unsupervised analysis so that topics will be ranked by semantic legitimacy. Compared to all the aforementioned approaches, techniques proposed in this paper not only performs topic evaluation but also directly filters junk contents thus effectively refining a topic by discovering beneficial contents with respect to a topic. The expected outcome is more coherent topic modeling consisting of less garbage and topic aggregation with higher quality.

7. Conclusion

The objective of this paper is to extract closely correlated documents and filter less relevant items from a set of already similar documents clustered by the LDA topic model. It first discussed the intuition of recognizing document relevant on themes and how it is achieved by manual labor. Next it designed an automatic framework to simulate the manual work. The major approach relies on the search engine suggests provided by the search engine as a critical characteristic to infer semantic themes contained in each document. To help enhance such expectation from search engine suggests it shows that inter-document similarity measures can verify the suggest frequency schemes and improve model output to certain extents depending on scenarios of different query focuses. Future work involves re-evaluating topic clusters after the proposed model is applied using topic coherence metrics mentioned in the related works.

(注5)：https://wordnet.princeton.edu/

References

[1] N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In *Proc. of 10th IWCS*, pages 13–22, 2013.

[2] L. AlSumait, D. Bardara, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *Proc. ECML/PKDD*, pages 67–82, 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] H. Chan and L. Akoglu. External evaluation of topic models: A graph mining approach. In *Proc. 13th ICDM*, pages 973–978, 2013.

[5] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Proc. NIPS*, pages 288–296, 2009.

[6] F. Chierichetti, R. Kumar, S. Pandey, and S. Vassilvitskii. Finding the Jaccard median. In *Proc. 21st ACM-SIAM Symposium on Discrete Algorithms*, pages 293–311, 2010.

[7] A. de Waal and E. Barnard. Evaluating topic models with stability. In *Proc. PRASA*, pages 79–84, 2008.

[8] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proc. 14th EACL*, pages 530–539, 2014.

[9] D. Mimno and D. Blei. Bayesian checking for topic models. In *Proc. EMNLP*, pages 227–237, 2011.

[10] C. C. Musat, J. Velcin, S. Trausan-Matu, and M.-A. Rizoiu. Improving topic evaluation using conceptual knowledge. In *Proc. 27th IJCAI*, pages 1866–1871, 2011.

[11] D. Newman, S. Karimi, and L. Cavedon. External evaluation of topic models. In *Proc. in Australasian Document Computing Symposium*, pages 11–18, 2009.

[12] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proc. HLT-NAACL*, pages 100–108, 2010.

[13] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proc. 8th WSDM*, pages 399–408, 2015.

[14] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *Proc. EMNLP/CoNLL*, pages 952–961, 2012.

[15] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proc. 26th ICML*, pages 1105–1112, 2009.

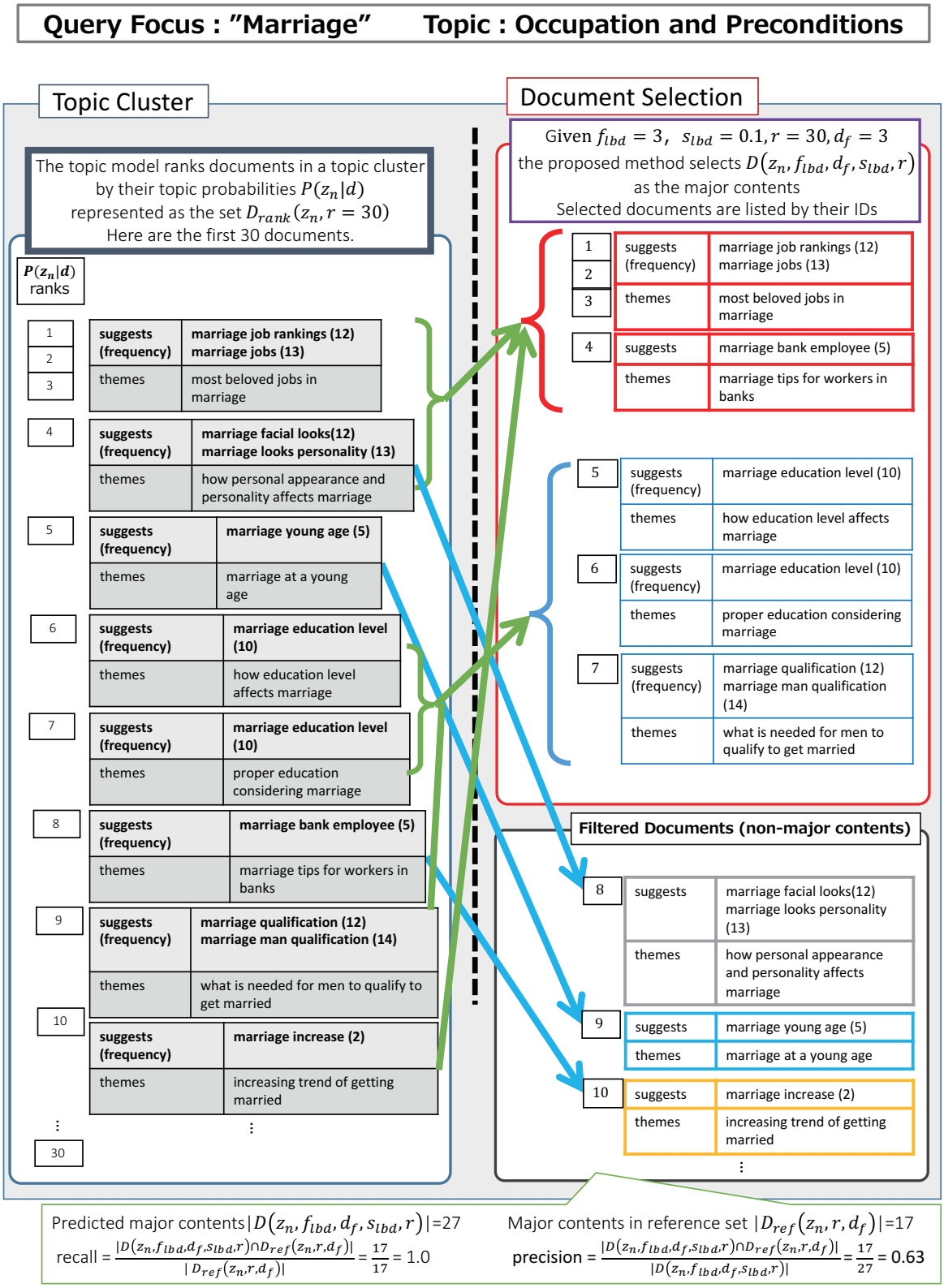

Figure 6 A Detailed Example of In-topic Major Contents Selection