Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016),
a Satellite Workshop of the NTCIR-12 Conference, June 7, 2016 Tokyo Japan

Two-layered Summaries for Mobile Search: Does the Evaluation Measure Reflect User Preferences?

Makoto P. Kato
Kyoto University
kato@dl.kuis.kyoto-u.ac.jp

Virgil Pavlu
Northeastern University
vip@ccs.neu.edu

Tetsuya Sakai
Waseda University
tetsuyasakai@acm.org

Takehiro Yamamoto
Kyoto University
tyamamot@dl.kuis.kyoto-u.ac.jp

Hajime Morita
Kyoto University
morita@nlp.ist.i.kyoto-u.ac.jp

ABSTRACT

This paper addresses two-layered summarization for mobile search, and proposes an evaluation framework for such summaries. A single summary is not always satisfactory for all variety of users with different intents, and mobile devices impose hard constraints on the summary format. In a two-layered summary, the first layer contains general useful information, while the second layer contains information interesting for different types of users. As users with different interests can take their own reading paths, they could find their desired information more efficiently than if all layers are presented as one block of text, by skipping certain parts of the second layer. Our proposed evaluation metric, *M-measure*, takes into account all the possible reading paths in a two-layered summary, and is defined as the expected utility of these paths. Our user study compared M-measure with pairwise user preferences on two-layered summaries, and found that M-measure agrees with the user preferences on more than 70% summary pairs.

1. INTRODUCTION

Web search engines usually return a ranked list of URLs in response to a query. After typing the query and clicking on the search button, the user often has to visit several Web pages and locate relevant parts within those pages. Especially for mobile users, these actions require significant effort and attention on a crowded small screen; they could be avoided if a system returned a concise summary of relevant information to the query. Such query-focused summarization techniques have been proposed and evaluated in NTCIR 1CLICK tasks [1, 7] and MobileClick tasks [2, 3], and recently gained attention due to the growth of mobile searchers.

In this paper, we focus on two-layered summarization, which allows users with diverse intents to effectively find their desired information, as opposed to read the block of text for a traditional single-layer summary. A two-layered summary consists of the first and second layers as shown in Figure 1. The first layer is expected to contain information interesting for most of the users, and the links to the second layer; the second layer, which is hidden until its header link is clicked on, is expected to contain information relevant for a particular set of users. In a two-layered summary, users can avoid reading text in which they are not interested, thus save time spent on non-relevant information, if they can make a binary yes/no decision of each second-layer entry from the head link alone. As an example, Wikipedia presents its pages in a two-layer format by default on mobile devices[1].

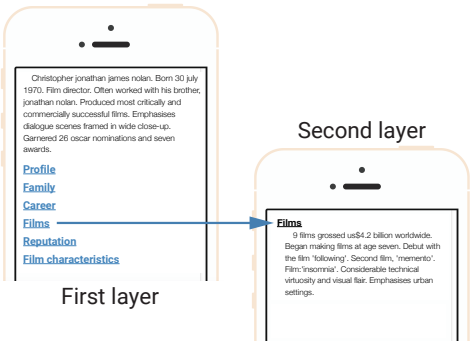


Figure 1: A two-layered summary for query "`christopher nolan`". Users can see the second layer if they click on a link in the first layer.

We propose an evaluation metric called *M-measure* designed for two-layered summaries. Although summarization evaluation has been studied extensively (*e.g.* ROUGE [4]), few studies apply to multi-layer format. One of the challenges in two-layered summarization evaluation is the uncertainty of a particular user *trailtexts*, *i.e.* possible paths of reading a two-layered summary; in general there are 2^N possible different paths, given N links. We can enumerate user models that go through different paths on a two-layered summary following each user interest, and generate multiple trailtexts. Then we compute the utility of each trailtext by *U-measure* [5], which is an evaluation metric for general purpose summarization. The M-measure is defined as the expected U-measure over the trailtexts, i.e. the sum of the trailtext utility weighted by the probability of the trailtext being read.

We ask how well M-measure reflects the user preferences on two-layered summaries. We run a user study designed to answer this question: Pairs of two-layered summaries were shown to assessors from crowd-sourcing services, and were evaluated by pairwise comparison. Our experimental results show that M-measure is in accord with the user preferences for more than 70% of the pairs. In addition, by comparing M-measure with its simpler variants, we argue that each component of M-measure is necessary for proper reflection of the user preferences.

[1] `https://en.wikipedia.org/wiki/Help:Mobile_access`

2. TWO-LAYERED SUMMARIZATION

In order to describe our setup for two-layered summarization, we first introduce two notions: *iUnits* and *intents*.

Information Units (or *iUnits*, used in NTCIR 1CLICK tasks [1, 7]) are the building blocks of any two-layered summary. iUnits are atomic pieces of important information for a query. For example, "`born 30 July 1970`", "`film director`", and "`debut with the film 'following'`" are iUnits for query "`christopher nolan`". iUnits should be *relevant*, i.e. provide useful factual information to the user on its own; and *atomic*, i.e. an iUnit cannot be broken down into multiple iUnits without loss of the original semantics.

The *intent* is a notion also used in the NTCIR INTENT task [10], and is used as the anchor text of links in a two-layered summary. An intent is textual representation of a certain topic in which users who input a particular query are interested. For example, "`career`" and "`reputation`" are intents for the query above.

Letting q be a query, U_q be a set of iUnits for q, and I_q be a set of intents for q, the formal problem of two-layered summarization is defined in this paper as follows: Given q, U_q, and I_q, generate a two-layer summary that consists of the first layer $\mathbf{f}$ and second layer $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. The first layer $\mathbf{f}$ consists of iUnits and links (*e.g.* $\mathbf{f} = (u_1, u_2, i_1, u_3)$ where $u_j \in U_q$ is an iUnit and $i_j \in I_q$ is a link/intent). Each link i_j links to the second layer $\mathbf{s}_j$, and must be one of the provided intents I_q. A second layer $\mathbf{s}_j$ is composed of only iUnits (*e.g.* $\mathbf{s}_1 = (u_{1,1}, u_{1,2}, u_{1,3})$). Note that this problem setting differs from traditional summarization problems in that all the relevant information, iUnits, are given. We use this setting to increase the reusability of test collections developed for the two-layered summarization.

3. EVALUATION METRIC

Intuitively, a two-layered summary is good if: (1) The summary does not include non-relevant iUnits in the first layer; (2) The first layer includes iUnits relevant for all the intents; and (3) iUnits in the second layer are relevant for the intent that links to them.

Our design for the evaluation metric makes the following choices and assumptions:

- Users are interested in one of the intents $i \in I_q$ by following the intent probability $P(i|q)$.
- Users read a summary following these rules:
  (1) Start at the beginning of the first layer.
  (2) When reaching the end of a link i_j which interests the users, click on the link and start to read its second layer $\mathbf{s}_j$.
  (3) When reaching the end of the second layer $\mathbf{s}_j$, go back to the end of the link i_j and continue reading.
  (4) Stop after reading no more than L characters.
- We choose as the base-measure for utility of text the U-measure proposed by Sakai and Dou [5], which consists of a position-based gain and a position-based decay function. We could choose a different summary evaluation base-measure (*e.g.* ROUGE [4]).
- The two-layer evaluation metric is the expected utility of text read by users.

We generate the user trailtexts according to the user model above, compute a U-measure score for each trailtext, and finally estimate the expected U-measure by combining all the U-measure scores of trailtexts. *M-measure*, an evaluation metric for the two-layered summarization, is defined as follows:

$$M = \sum_{\mathbf{t} \in T} P(\mathbf{t}) U(\mathbf{t}), \tag{1}$$

where T is a set of all possible trailtexts, $P(\mathbf{t})$ is a probability of going through a trailtext $\mathbf{t}$, and $U(\mathbf{t})$ is the U-measure score of a trailtext $\mathbf{t}$.

A trailtext is a concatenation of all the texts read by a user, and can be defined as a list of iUnits and links in our case. According to our user model, a trailtext of a user who is interested in intent i can be obtained by inserting a list of iUnits in the second layer $\mathbf{s}_j$ after the link of i_j. More specifically, given the first layer $\mathbf{f} = (u_1, \ldots, u_{j-1}, i_k, u_j, \ldots)$ and second layer $\mathbf{s}_k = (u_{k,1}, \ldots, u_{k,|\mathbf{s}_k|})$, trailtext $\mathbf{t}_{i_k}$ of intent i_k is defined as follows: $\mathbf{t}_{i_k} = (u_1, \ldots, u_{j-1}, i_k, u_{k,1}, \ldots, u_{k,|\mathbf{s}_k|}, u_j, \ldots)$.

We consider only the trailtexts that correspond to users intents, thus the probability of a trailtext is equivalent to that of the intent for which the trailtext is generated. Then the M-measure can be written as

$$M = \sum_{i \in I_q} P(i|q) U_i(\mathbf{t}_i). \tag{2}$$

where the base-measure U is now measured in terms of intent i in the equation above, since we assume that users going through $\mathbf{t}_i$ are interested in intent i.

The computation of U-measure [5] involves the importance and offset of each relevant iUnits in a trailtext. The offset of iUnit u is defined as the number of characters between the beginning of the trailtext and the end of u. More precisely, the offset of the j-th iUnit in trailtext $\mathbf{t}$ is $\mathrm{pos}_{\mathbf{t}}(u_j) = \sum_{j'=1}^{j} \mathrm{chars}(u_{j'})$, where $\mathrm{chars}(u)$ is the number of characters of iUnit u except symbols and white spaces. Note that a link in the trailtext is regarded as a non-relevant iUnit for the sake of convenience. U-measure is defined as follows:

$$U_i(\mathbf{t}) = \frac{1}{\mathcal{N}} \sum_{j=1}^{|\mathbf{t}|} g_i(u_j) d(u_j), \tag{3}$$

where $g_i(u_j)$ is the importance of iUnit u_j in terms of intent i, d is a position-based decay function, and $\mathcal{N}$ is a normalization factor (we set $\mathcal{N}$=1). The position-based decay function used is $d(u) = \max\left(0, 1 - \frac{\mathrm{pos}_{\mathbf{t}}(u)}{L}\right)$, where L is a patience parameter of users. Note that no gain can be obtained after L characters read, *i.e.* $d(u) = 0$. This is consistent with our user model in which users stop after reading L characters.

4. EXPERIMENTS

In this section, we describe data used for computing M-measure, explain pairwise comparison of two-layered summaries, and show results of comparison of M-measure and user preferences. In the following experiments, two crowd-sourcing services were used: CrowdFlower[2] (English) and Lancers[3] (Japanese).

4.1 Data

We used a test collection provided by NTCIR-12 MobileClick-2 [3], and evaluated system results submitted to this evaluation campaign. The test collection contains 100 English and 100 Japanese queries, a set of iUnits manually created for each query (23.8 per English query, and 41.7 per Japanese query), and a set

[2] `http://www.crowdflower.com/`
[3] `http://www.lancers.jp/`

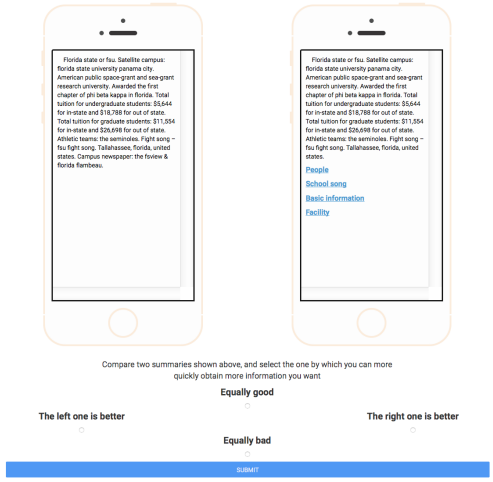


Figure 2: Pairwise comparison interface.

of intents manually created for each query (4.48 per English query, and 4.37 per Japanese query).

The intent probability $P(i|q)$ was obtained by votes from 10 crowd-sourcing workers. Workers were asked to vote for all the intents they were interested in, with respect to that query. We normalized the number of votes for each intent, *i.e.* $P(i|q) = n_{i,q}/n_{\cdot,q}$ where $n_{i,q}$ is the number of votes intent i received, and $n_{\cdot,q}$ is the total number of votes for query q.

The importance of an iUnit in terms of a certain intent, $g_i(u_j)$, was evaluated at a five-point scale: 0 (unimportant), 1, 2 (somewhat important), 3, and 4 (highly important). For example, in response to query q="`yosemite`", iUnit "`located in California`" is *unimportant* for intent "`Mac OS`", while it is *highly important* for intent "`US national park`". Two assessors were instructed to evaluate each iUnit importance, explicitly assuming interest in the corresponding intent. The average assessors' importance score was used for evaluation. The inter-assessor agreement was moderate, 0.556 in terms of *quadratic-weighted kappa* [9].

4.2 Pairwise Comparison

We showed pairs of two-layered summaries to workers in crowd-sourcing services, and asked them to judge which summary is better. The crowd-sourcing procedure is summarized as follows:

(1) Showed a list of queries and let workers select the one in which they are the most interested,
(2) Asked the workers to search for basic information about the query for at least three minutes,
(3) Showed a pair of two-layered summaries as shown in Figure 2 and let the workers select from the following options: *the left one is better*, *the right one is better*, *equally good*, and *equally bad*.

We allowed the workers to select the most interesting query so that they could judge two-layered summaries from the viewpoint of users actually interested in such a query. Search was required in order to ensure a minimum and uniform query familiarization. The pairwise comparison criteria shown to the workers were (1) how much useful information you can get from the summary, and (2) how quickly you can get useful information from the summary.

In this experiment, we used the 25 out of 100 most frequent queries, and the respective summaries of seven systems for each of English and Japanese. The query frequency was estimated by using Google AdWords Keyword Planner[4]. For each query, we tasked each worker with all the pairs of the seven systems, including some repeating *validation pairs* allowing us to check worker consistency. We excluded workers who (1) did not spend 200s for search on average, or (2) did not give consistent answers to at least 40% of the validation pairs. On average, 14 workers were hired per query, and they were paid $150 and 200JPY in English and Japanese tasks, respectively. As a result, we obtained pairwise preferences for $25 * (7 * 6/2) = 525$ pairs of summaries.

4.3 Results

Each dot in Figures 3 and 4 represents a pair of systems (R, R') for a particular query, with coordinates the difference in terms of M-measure ($M(R) - M(R')$) on x-axis, and $y_{RR'}$ the fraction of preferences to R over R' on y-axis. Judgements *equally good* and *equally bad* were regarded as 0.5 votes for both systems for y-axis calculation. We expected that users votes for R if $M(R) - M(R')$ is positive (*i.e.* $M(R) > M(R')$), while users votes for R' if $M(R) - M(R')$ is negative (*i.e.* $M(R) < M(R')$), thus M-measure reflects the user preferences accurately if most of the dots are in the first and third quadrants (highlighted by yellow). Agreement in the figures is the fraction of system pairs, for all queries, where M-measure and user preferences agreed, *i.e.* $M(R) > M(R') \wedge y_{RR'} > 0.5$ or $M(R) < M(R') \wedge y_{RR'} < 0.5$. The agreement measure has been used to show how well evaluation metrics reflect user preferences [8], as opposed to the Kendall's τ usually used when full ranked lists are available (our system pairs for each plot are per-query pairs, for all queries).

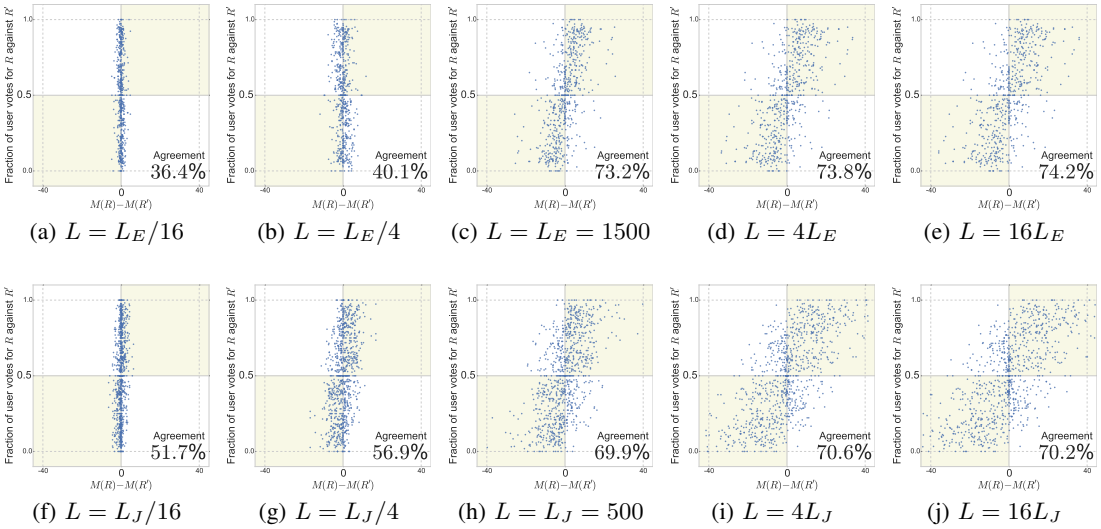
Figure 3 shows the effect of the parameter L in U-measure. Recall that L represents the user patience and controls the position-based decay. The larger L is, the smaller the decay is, *i.e.* users in our model are less sensitive to the position of iUnits. The default value for L was $L_E = 1,500$ for English and $L_J = 500$ for Japanese, as $L_J = 500$ was recommended [6], and the average character reading speed for English is approximately three times as fast as that for Japanese [11]. The results showed that large L could better reflect the user preferences: The highest agreement achieved was 74.2% by (e) $16L_E$ for English, and 70.6% by (i) $4L_J$ for Japanese. Note that the crowd-sourcing workers were possibly more patient than searchers: they spent 26.1 and 22.5 seconds for each pairwise comparison. Therefore, L could be further adjusted by comparing the assessment time by the workers and searchers.

Figure 4 shows the results of simplified versions of M-measure. *Only first layer* (b and f) is the result when only iUnits in the first layer were evaluated, while *only second layer* (c and g) is the result when only iUnits in the second layer were evaluated. The agreement was quite low if only the first layer was used, whereas the agreement of M-measure with only the second layer was close to that of the original M-measure. This suggests that the preference of the second layer highly affected the overall preference, or the quality of the second layer highly correlates to the overall quality. *Uniform* $P(i|q)$ (d and h) is the result when $P(i|q)$ was a uniform distribution, and achieved slightly lower agreement than the other plots where M-measure used the intent probability estimated by user votes. This result suggests that it is important to take into account the intent probability in evaluation.

5. CONCLUSIONS

This paper addressed two-layered summarization for mobile search, and proposed an evaluation framework for such summaries. We compared M-measure with pairwise user preferences on two-

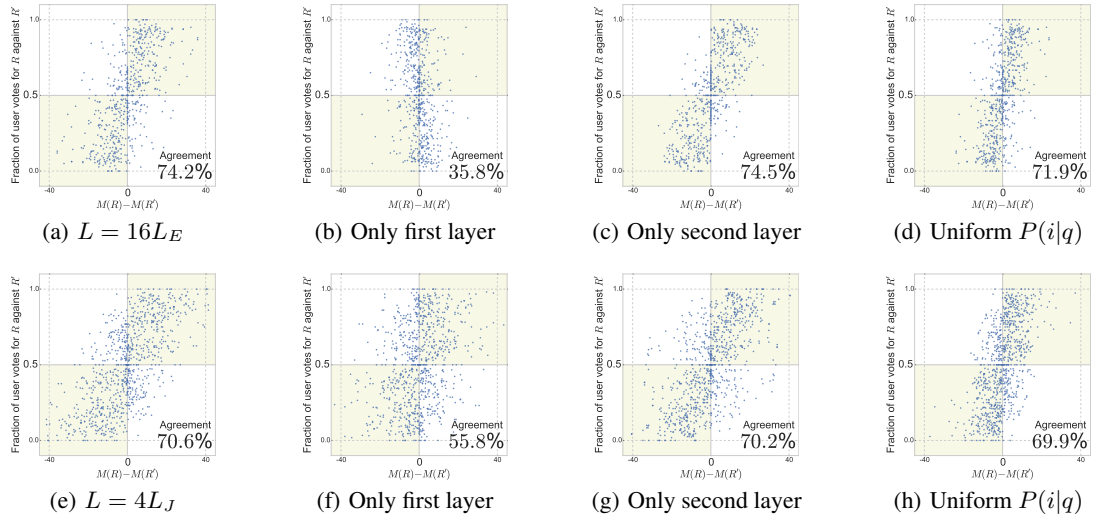
[4] `https://adwords.google.com/KeywordPlanner`



(a) $L = L_E/16$ (b) $L = L_E/4$ (c) $L = L_E = 1500$ (d) $L = 4L_E$ (e) $L = 16L_E$

(f) $L = L_J/16$ (g) $L = L_J/4$ (h) $L = L_J = 500$ (i) $L = 4L_J$ (j) $L = 16L_J$

Figure 3: Pairwise user preferences vs. difference of summary pairs in terms of M-measure with different values for L; English (top) and Japanese (bottom).



(a) $L = 16L_E$ (b) Only first layer (c) Only second layer (d) Uniform $P(i|q)$

(e) $L = 4L_J$ (f) Only first layer (g) Only second layer (h) Uniform $P(i|q)$

Figure 4: Pairwise user preferences vs. M-measure difference of summary pairs; English (top) and Japanese (bottom).

layered summaries. In summary, we found that 1) there was over 70% agreement between M-measure and user preferences, 2) M-measure with a larger L (*i.e.* more user patience) than that recommended in the previous work [6] could achieve higher agreement with the user preferences, and 3) an evaluation metric for two-layered summaries should take into account the second layer and intent probability for better reflection of the user preferences.

6. REFERENCES

[1] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the NTCIR-10 1CLICK-2 Task. In *NTCIR-10 Conference*, pages 243–249, 2013.

[2] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the NTCIR-11 MobileClick Task. In *NTCIR-11 Conference*, pages 195–207, 2014.

[3] M. P. Kato, T. Sakai, T. Yamamoto, V. Pavlu, H. Morita, and S. Fujita. Overview of the NTCIR-12 MobileClick-2 Task. In *NTCIR-12 Conference*, 2016.

[4] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004.

[5] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In *SIGIR*, pages 473–482, 2013.

[6] T. Sakai and M. P. Kato. One click one revisited: Enhancing evaluation based on information units. In *AIRS*, pages 39–51, 2012.

[7] T. Sakai, M. P. Kato, and Y.-I. Song. Overview of NTCIR-9 1CLICK. In *NTCIR-9*, pages 180–201, 2011.

[8] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR*, pages 555–562, 2010.

[9] J. Sim and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

[10] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *NTCIR-9*, pages 82–105, 2011.

[11] S. Trauzettel-Klosinski and K. Dietz. Standardized assessment of reading performance: The new international reading speed texts IReST. *Investigative Ophthalmology & Visual Science*, 53(9):5452–5461, 2012.