Summarize What You Are Interested In: An Optimization Framework for Interactive Personalized Summarization

Rui Yan
Department of Computer
Science and Technology,
Peking University,
Beijing 100871, China
r.yan@pku.edu.cn

Jian-Yun Nie
Département d´informatique
et de recherche opérationnelle,
Université de Montréal,
Montréal, H3C 3J7 Québec, Canada
nie@iro.umontreal.ca

Xiaoming Li
Department of Computer
Science and Technology,
Peking University,
Beijing 100871, China
lxm@pku.edu.cn

Abstract

Most traditional summarization methods treat their outputs as static and plain texts, which fail to capture user interests during summarization because the generated summaries are the same for different users. However, users have individual preferences on a particular source document collection and obviously a universal summary for all users might not always be satisfactory. Hence we investigate an important and challenging problem in summary generation, i.e., Interactive Personalized Summarization (IPS), which generates summaries in an interactive and personalized manner. Given the source documents, IPS captures user interests by enabling interactive clicks and incorporates personalization by modeling captured reader preference. We develop experimental systems to compare 5 rival algorithms on 4 instinctively different datasets which amount to 5197 documents. Evaluation results in ROUGE metrics indicate the comparable performance between IPS and the best competing system but IPS produces summaries with much more user satisfaction according to evaluator ratings. Besides, low ROUGE consistency among these user preferred summaries indicates the existence of personalization.

1 Introduction

In the era of information explosion, people need new information to update their knowledge whilst information on Web is updating extremely fast. Multi-document summarization has been proposed to address such dilemma by producing a summary delivering the majority of information content from a document set, and hence is a necessity.

Traditional summarization methods play an important role with the exponential document growth on the Web. However, for the readers, the impact of human interests has seldom been considered. Traditional summarization utilizes the same methodology to generate the same summary no matter who is reading. However, users may have bias on what they prefer to read due to their potential interests: they need *personalization*. Therefore, traditional summarization methods are to some extent insufficient.

Topic biased summarization tries for personalization by pre-defining human interests as several general categories, such as *health* or *science*. Readers are required to select their possible interests before summary generation so that the chosen topic has priority during summarization. Unfortunately, such topic biased summarization is not sufficient for two reasons: (1) interests cannot usually be accurately pre-defined by ambiguous topic categories and (2) user interests cannot always be foreknown. Often users do not really know what general ideas or detail information they are interested in until they read the summaries. Therefore, more flexible *interactions* are required to establish personalization.

Due to all the insufficiencies of existed summarization approaches, we introduce a new multi-document summarization task of Interactive Personalized Summarization (IPS) and a novel solution for the task. Taking a document collection as input, the system outputs a summary aligned both with source corpus and with user personalization, which is captured by flexible *human−system* interactions. We

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1342–1351,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

build an experimental system on 4 real datasets to verify the effectiveness of our methods compared with 4 rivals. The contribution of IPS is manifold by addressing following challenges:

- The **1st challenge** for IPS is to integrate user interests into traditional summary components. We measure the utilities of these components and combine them. We formulate the task into a balanced optimization framework via iterative substitution to generate summaries with maximum overall utilities.
- The **2nd challenge** is to capture user interests through interaction. We develop an interactive mechanism of "click" and "examine" between readers and summaries and address sparse data by "click smoothing" under the scenario of few user clicks.

We start by reviewing previous works. In Section 3 we provide IPS overview, describe user interaction and optimize component combination with personalization. We conduct empirical evaluation and demonstrate the experimental system in Section 4. Finally we draw conclusions in Section 5.

2 Related Work

Multi-Document Summarization (MDS) has drawn much attention in recent years and gained emphasis in conferences such as ACL, EMNLP and SIGIR, etc. General MDS can either be extractive or abstractive. The former assigns salient scores to semantic units (e.g. sentences, paragraphs) of the documents indicating their importance and then extracts top ranked ones, while the latter demands information fusion(e.g. sentence compression and reformulation). Here we focus on extractive summarization.

Centroid-based method is one of the most popular extractive summarization method. MEAD (Radev et al., 2004) and NeATS (Lin and Hovy, 2002) are such implementations, using position and term frequency, etc. MMR (Goldstein et al., 1999) algorithm is used to remove redundancy. Most recently, the graph-based ranking methods have been proposed to rank sentences or passages based on the "votes" or "recommendations" between each other. The graph-based methods first construct a graph representing the sentence relationships at different granularities and then evaluate the saliency score of the sentences based on the graph. TextRank (Mihalcea and Tarau, 2005) and LexPageRank (Erkan and Radev, 2004) use algorithms similar to PageRank and HITS to compute sentence importance. Wan et al. improve the graph-ranking algorithm by differentiating intra-document and inter-document links between sentences (2007b) and incorporate cluster information in the graph model to evaluate sentences (2008).

To date, topics (or themes, clusters) in documents have been discovered and used for sentence selection for topic biased summarization (Wan and Yang, 2008; Gong and Liu, 2001). Wan et al. have proposed a manifold-ranking method to make uniform use of sentence-to-sentence and sentence-to-topic relationships to generate topic biased summaries (2007a). Leuski et al. in (2003) pre-define several topic concepts, assuming users will foresee their interested topics and then generate the topic biased summary. However, such assumption is not quite reasonable because user interests may not be forecasted, or pre-defined accurately as we have explained in last section.

The above algorithms are usually traditional extensions of generic summarizers. They do not involve interactive mechanisms to capture reader interests, nor do they utilize user preference for personalization in summarization. Wan et al. in (2008) have proposed a summarization biased to neighboring reading context through anchor texts. However, such scenario does not apply to contexts without human-edited anchor texts like Wikipedia they have used. Our approach can naturally and simultaneously take into account traditional summary elements and user interests and combine both in optimization under a wider practical scenario.

3 Interactive Personalized Summarization

Personalization based on user preference can be captured via various alternative ways, such as *eye-tracking* or *mouse-tracking* instruments used in (Guo and Agichtein, 2010). In this study, we utilize interactive user clicks/examinations for personalization.

Unlike traditional summarization, IPS supports *human−system* interaction by *clicking* into the summary sentences and *examining* source contexts. The implicit feedback of user clicks indicates what they are interested in and the system collects preference information to update summaries if readers wish to. We obtain an associated tuple $<q, c>$ between a clicked sentence **q** and the examined contexts **c**.

As q has close semantic coherence with neighboring contexts due to consistency in human natural language, we consider a window of sentences centered at the clicked sentence q as c, which is a bag of sentences. The window size k is a parameter to set.

However, click data is often sparse: users are not likely to click more than 1/10 of total summary sentences within a single generation. We amplify these tiny hints of user interest by *click smoothing*.

We change the flat summary structure into a hierarchical organization by extracting important semantic units (denoted as **u**) and establishing linkage between them. If the clicked sentence q contains u, we diffuse the click impact to the correlated units, which makes a single click perform as multiple clicks and the sparse data is smoothed.

Problem Formulation

Input: Given the sentence collection D decomposed by documents, $D = \{s_1, s_2, \dots, s_{|D|}\}$ and the clicked sentence record $Q = \{q_1, q_2, \dots\}$, we generate summaries in sentences. A user click is associated with a tuple $<q, (u), c>$ where the existence of u depends on whether q contains u. The collection of semantic units is denoted as $M = \{u_1, u_2, \dots, u_{|M|}\}$.

Output: A summary S as a set of sentences $\{s_1, s_2, \dots, s_{|S|}\}$ and $S \subset D$ according to the pre-specified compression rate ϕ ($0 < \phi < 1$).

After the overview and formulation of IPS problem, we move on to the major components of *User Interaction* and *Personalized Summarization*.

3.1 User Interaction

Hypertexify Summaries. We hypertexify the summary structure by establishing linkage between semantic units. There are several possible formats for semantic units, such as words or n-grams, etc. As single words are proved to be not illustrative of semantic meanings (Zhao et al., 2011) and n-grams are rigid in length, we choose to extract semantic units at a phrase granularity. Among all phrases from source texts, some are of higher importance to attract user interests, such as hot concepts or popular event names. We utilize the toolkit provided by (Zhao et al., 2011) based on graph proximity LDA (Blei et al., 2003) to extract key phrases and their corresponding topic. A topic T is represented by $\{(u_1, \pi(u_1, T)), (u_2, \pi(u_2, T)), \dots\}$ where $\pi(u, T)$ is the probability of u belonging to topic T. We invert the topic-unit representation in Table 1, where each u is represented as a topic vector. The correlation $corr(.)$ between u_i, u_j is measured by cosine similarity $sim(.)$ on topic distribution vector $\vec{u}_i$, $\vec{u}_j$.

$$corr(u_i, u_j) = sim_{\text{topic}}(\vec{u}_i, \vec{u}_j) \qquad (1)$$

Table 1: Inverted representation of topic-unit vector.

$\vec{u}_1$	$\pi(u_1, T_1)$	$\pi(u_1, T_2)$	$\dots$	$\pi(u_1, T_n)$
$\vec{u}_2$	$\pi(u_2, T_1)$	$\pi(u_2, T_2)$	$\dots$	$\pi(u_2, T_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vec{u}_{\|M\|}$	$\pi(u_{\|M\|}, T_1)$	$\pi(u_{\|M\|}, T_2)$	$\dots$	$\pi(u_{\|M\|}, T_n)$

When the summary is *hypertexified* by established linkage, users click into the generated summary to examine what they are interested in. A single click on one sentence become multiple clicks via click smoothing when the indicative function $I(u|q) = 1$.

$$I(u|q) = \begin{cases} 1 & q \text{ contains } u; \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

The click smoothing brings pseudo clicks q' associated with u' and contexts c'. The entire user feedback texts $\mathcal{A}$ from q can be written as:

$$\mathcal{A}(q) = I(u|q) \sum_{j=1}^{|M|} corr(u', u)(u' + \gamma \cdot c') + \gamma \cdot c \quad (3)$$

where γ is the weight tradeoff between u and associated contexts c. If $I(u|q) = 0$, only the examined context c is feedbacked for user preference; otherwise, correlative contexts with u are taken into consideration, which is a process of impact diffusion.

3.2 Personalized Summarization

Traditional summarization involves two essential requirements: (1) **coverage**: the summary should keep alignment with the source collection, which is proved to be significant (Li et al., 2009). (2) **diversity**: according to MMR principle (Goldstein et al., 1999) and its applications (Wan et al., 2007b; Wan and Yang, 2008), a good summary should be concise and contain as few redundant sentences as possible, i.e., two sentences providing similar information should not both present. According to our investigation, we observe that a well generated summary should properly consider a key component of (3) **user interests**, which captures user preference to summarize what they are interested in.

All above requirements involve a measurement of similarity between two word distributions Θ_1 and Θ_2. Cosine, Kullback-Leibler divergence D_{KL} and Jensen Shannon divergence D_{JS} are all able to measure the similarity, but (Louis and Nenkova, 2009) indicate the superiority of D_{JS} in summarization task. We also introduce a pair of decreasing/increasing logistic functions, $\mathcal{L}_1(x) = 1/(1+e^x)$ and $\mathcal{L}_2(x) = e^x/(1+e^x)$, to map the divergence into interval [0,1]. V is the vocabulary set and *tf* denotes the term frequency for word w.

$$D_{JS}(\Theta_1||\Theta_2) = \frac{1}{2}[D_{KL}(\Theta_1||\Theta_2)+D_{KL}(\Theta_2||\Theta_1)]$$

where

$$D_{KL}(\Theta_1||\Theta_2) = \sum_{k\in V} p(w|\Theta_1)log\frac{p(w|\Theta_1)}{p(w|\Theta_2)}$$

where

$$p(w|\Theta) = \frac{tf(w,\Theta)}{\sum_{w'} tf(w',\Theta)}.$$

Modeling Interest for User Utility. Given a generated summary S, users tend to scrutinize texts relevant to their interests. Texts related to user implicit feedback are collected as $\mathcal{A} = \sum_{i=1}^{|Q|} \mathcal{A}(q_i)$. Intuitively, the smaller distance between the word distribution of final summary (Θ_S) and the word distribution of user preference ($\Theta_{\mathcal{A}}$), the higher utility of user interests $\mathcal{U}_{user}(S)$ will be, i.e.,

$$\mathcal{U}_{user}(S) = \mathcal{L}_1(D_{JS}(\Theta_S||\Theta_{\mathcal{A}})). \qquad (4)$$

We model the utility of traditional summarization $\mathcal{U}_{trad}(S)$ using a linear interpolation controlled by parameter δ between utility from *coverage* $\mathcal{U}_c(S)$ and utility $\mathcal{U}_d(S)$ from *diversity*:

$$\mathcal{U}_{trad}(S) = \mathcal{U}_c(S) + \delta\cdot\mathcal{U}_d(S). \qquad (5)$$

Coverage Utility. The summary should share a closer word distribution with the source collection (Allan et al., 2001; Li et al., 2009). A good summary focuses on minimizing the loss of main information from the whole collection D. Utility from coverage $\mathcal{U}_c(S)$ is defined as follows and for coverage utility, smaller divergence is desired.

$$\mathcal{U}_c(S) = \mathcal{L}_1(D_{JS}(\Theta_S||\Theta_D)). \qquad (6)$$

Diversity Utility. Diversity measures the novelty degree of any sentence s compared with all other sentences within S, i.e., the distances between all other sentences and itself. Diversity utility $\mathcal{U}_d(S)$ is an average novelty score for all sentences in S. For diversity utility, larger distance is desired, and hence we use the increasing function $\mathcal{L}_2$ as follows:

$$\mathcal{U}_d(S) = \frac{1}{|S|}\sum_{s\in S}\mathcal{L}_2(D_{JS}(\Theta_s||\Theta_{(S-s)})). \qquad (7)$$

3.3 Balanced Optimization Framework

A well generated summary S should be sufficiently aligned with the original source corpus, and also be optimized given the user interests. The utility of an individual summary $\mathcal{U}(S)$ is evaluated by the weighted combination of these components, controlled by parameter λ for balanced weights.

$$\mathcal{U}(S) = \mathcal{U}_{trad}(S) + \lambda\cdot\mathcal{U}_{user}(S) \qquad (8)$$

Given the sentence set D and the compression rate ϕ, there are $\phi\cdot|D|$ out of $|D|$ possibilities to generate S. The IPS task is to predict the optimized sentence subset of S^* from the space of all combinations. The objective function is as follows:

$$S^* = \underset{S}{\operatorname{argmax}}\,\mathcal{U}(S). \qquad (9)$$

As $\mathcal{U}(S)$ is measured based on preferred interests from user interaction within a generation in our system, we extract S iteratively to approximate S^*, i.e, maximize $\mathcal{U}(S)$ based on the user feedbacks from the interaction sessions. Each session is an iteration. We use a similar framework as we have proposed in (Yan et al., 2011).

During every session, the top ranked sentences are strong candidates for the summary to generate and the rank methodology is based on the metrics $\mathcal{U}(.)$. The algorithm tends to highly rank sentences which are with both coverage utility and interest utility, and are diversified in balance: we rank each sentence s according to $\mathcal{U}(s)$ under such metrics.

Consider $S^{(n-1)}$ generated in the (*n-1*)-th session which consists of top $\phi|D|$ ranked sentences, as well as the top $\phi|D|$ ranked sentences in the n-th iteration (denoted by $\mathcal{O}^{(n)}$), they have an intersection set of $\mathcal{Z}^{(n)} = S^{n-1} \cap \mathcal{O}^{n}$. There is a substitutable sentence set $\mathcal{X}^{(n)} = S^{(n-1)} - \mathcal{Z}^{(n)}$ and a new candidate sentence set $\mathcal{Y}^{(n)} = \mathcal{O}^{(n)} - \mathcal{Z}^{(n)}$. We substitute $\mathrm{x}^{(n)}$ sentences with $\mathrm{y}^{(n)}$, where $\mathrm{x}^{(n)} \subseteq \mathcal{X}^{(n)}$ and $\mathrm{y}^{(n)} \subseteq \mathcal{Y}^{(n)}$. During every iteration, our goal is to find a substitutive pair $<\mathrm{x}, \mathrm{y}>$ for S:

$$<\mathrm{x}, \mathrm{y}> : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{R}.$$

To measure the performance of such a substitution, a discriminant utility gain function $\Delta\mathcal{U}_{\mathrm{x},\mathrm{y}}$

$$\begin{aligned} \Delta\mathcal{U}^{(n)}_{\mathrm{x}^{(n)},\mathrm{y}^{(n)}} &= \mathcal{U}(S^{(n)}) - \mathcal{U}(S^{(n-1)}) \\ &= \mathcal{U}((S^{(n-1)} - \mathrm{x}^{(n)}) \cup \mathrm{y}^{(n)}) - \mathcal{U}(S^{(n-1)}) \end{aligned} \tag{10}$$

is employed to quantify the penalty. Therefore, we predict the substitutive pair by maximizing the gain function $\Delta\mathcal{U}_{\mathrm{x},\mathrm{y}}$ over the state set $\mathcal{R}$, with a size of $\sum_{k=0}^{\mathcal{Y}} A^{k}_{\mathcal{X}} C^{k}_{\mathcal{Y}}$, where $<\mathrm{x}, \mathrm{y}> \in \mathcal{R}$. Finally the objective function of Equation (9) changes into maximization of utility gain by substitute $\hat{\mathrm{x}}$ with $\hat{\mathrm{y}}$ during each iteration:

$$< \hat{\mathrm{x}}, \hat{\mathrm{y}} >= \underset{\mathrm{x}\subseteq\mathcal{X},\mathrm{y}\subseteq\mathcal{Y}}{\mathrm{argmax}}\ \Delta\mathcal{U}_{\mathrm{x},\mathrm{y}}. \tag{11}$$

Note that the objectives of interest utility optimization and traditional utility optimization are not always the same because the word distributions in these texts are usually different. The substitutive pair $<\mathrm{x}, \mathrm{y}>$ may perform well based on the user preference component while not on the traditional summary part and vice versa. There is a tradeoff between both user optimization and traditional optimization and hence we need to balance them by λ.

The objective Equation (11) is actually to maximize $\Delta\mathcal{U}(S)$ from all possible substitutive pairs between two iteration sessions to generate S. The algorithm is shown in Algorithm 1. The threshold ϵ is set at 0.001 in this study.

Algorithm 1 Regenerative Optimization

```
Input: $D$, $\epsilon$, $\phi$
for all $s \in D$ do
    calculate $\mathcal{U}_{trad}(s)$
end for
$S \leftarrow$ top $\phi|D|$ ranked sentences
while new generation=TRUE do
    collect clicks and update utility from $\mathcal{U}'$ to $\mathcal{U}$
    if $|\mathcal{U}(S) - \mathcal{U}'(S)| > \epsilon$ then
        for all $s \in D$ do
            calculate $\mathcal{U}(s)$
        end for
        $\mathcal{O} \leftarrow$ top $\phi|D|$ ranked sentences by $\mathcal{U}(s)$
        $\mathcal{Z} \leftarrow S \cap \mathcal{O}$
        $\mathcal{X} \leftarrow S - \mathcal{Z}, \mathcal{Y} \leftarrow \mathcal{O} - \mathcal{Z}$
        for all $<\mathrm{x}, \mathrm{y}>$ pair where $\mathrm{x} \subseteq \mathcal{X}, \mathrm{y} \subseteq \mathcal{Y}$ do
            $\Delta\mathcal{U}_{\mathrm{x},\mathrm{y}} = \mathcal{U}((S - \mathrm{x}) \cup \mathrm{y}) - \mathcal{U}(S)$
        end for
        $< \hat{\mathrm{x}}, \hat{\mathrm{y}} >= \mathrm{argmax}\ \Delta\mathcal{U}_{\mathrm{x},\mathrm{y}}$
        $S \leftarrow (S - \hat{\mathrm{x}}) \cup \hat{\mathrm{y}}$
    end if
end while
```

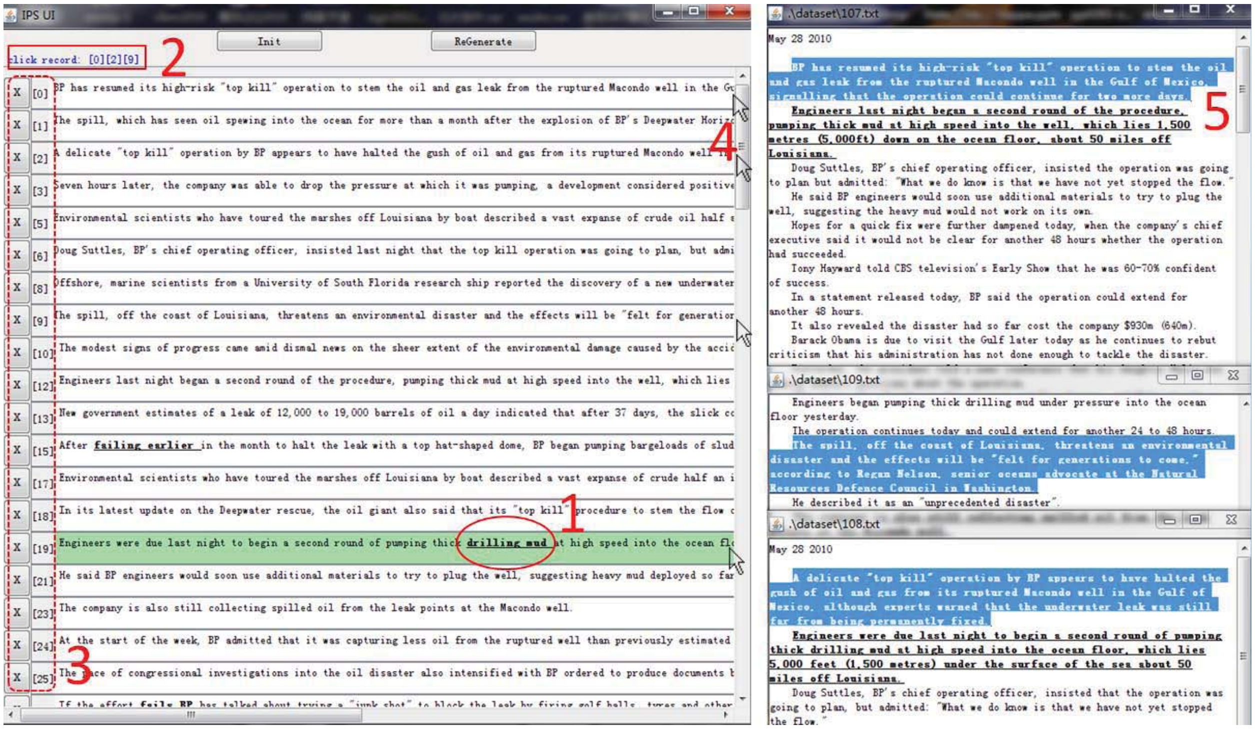
4 Experiments and Evaluation

4.1 Datasets

IPS can be tested on any document set but a tiny corpus to summarize may not cover abundant effective interests to attract user clicks indicating their preference. Besides, the scenario of small corpus is not quite practical for the exponential growing web. Therefore, we test IPS on large real world datasets. We build 4 news story sets which consist of documents and reference summaries to evaluate our proposed framework empirically. We downloaded 5197 news articles from 10 selected sources. As shown in Table 2, three of the sources are in UK, one of them is in China and the rest are in US. We choose them because many of these websites provide handcrafted summaries for their special reports, which serve as reference summaries. These events belong to different categories of Rule of Interpretation (ROI) (Kumaran and Allan, 2004). Statistics are in Table 3.

4.2 Experimental System Setups

- **Preprocessing.** Given a collection of documents, we first decompose them into sentences. Stop-words are removed and words stemming is performed. Then the word distributions can be calculated.
- **User Interface Design.** Users are required to specify the overall compression rate ϕ and the system extracts $\phi|D|$ sentences according to user utility



click record: [0][2][9]

X [0] BP has resumed its high-risk "top kill" operation to stem the oil and gas leak from the ruptured Macondo well in the Gu
X [1] The spill, which has seen oil spewing into the ocean for more than a month after the explosion of BP's Deepwater Horiz
X [2] A delicate "top kill" operation by BP appears to have halted the gush of oil and gas from its ruptured Macondo well in
X [3] Seven hours later, the company was able to drop the pressure at which it was pumping, a development considered positive
X [5] Environmental scientists who have toured the marshes off Louisiana by boat described a vast expanse of crude oil half a
X [6] Doug Suttles, BP's chief operating officer, insisted last night that the top kill operation was going to plan, but admi
X [8] Offshore, marine scientists from a University of South Florida research ship reported the discovery of a new underwater
X [9] The spill, off the coast of Louisiana, threatens an environmental disaster and the effects will be "felt for generation
X [10] The modest signs of progress came amid dismal news on the sheer extent of the environmental damage caused by the accic
X [12] Engineers last night began a second round of the procedure, pumping thick mud at high speed into the well, which lies
X [13] New government estimates of a leak of 12,000 to 19,000 barrels of oil a day indicated that after 37 days, the slick cc
X [15] After **failing earlier** in the month to halt the leak with a top hat-shaped dome, BP began pumping bargeloads of slud
X [17] Environmental scientists who have toured the marshes off Louisiana by boat described a vast expanse of crude half an i
X [18] In its latest update on the Deepwater rescue, the oil giant also said that its "top kill" procedure to stem the flow c
X [19] Engineers were due last night to begin a second round of pumping thick **drilling mud** at high speed into the ocean flc
X [21] He said BP engineers would soon use additional materials to try to plug the well, suggesting heavy mud deployed so far
X [23] The company is also still collecting spilled oil from the leak points at the Macondo well.
X [24] At the start of the week, BP admitted that it was capturing less oil from the ruptured well than previously estimated
X [25] The pace of congressional investigations into the oil disaster also intensified with BP ordered to produce documents t
If the effort **fails** BP has talked about trying a "junk shot" to block the leak by firing golf balls, tyres and other

May 28 2010

BP has resumed its high-risk "top kill" operation to stem the oil and gas leak from the ruptured Macondo well in the Gulf of Mexico, signalling that the operation could continue for two more days.

Engineers last night began a second round of the procedure, pumping thick mud at high speed into the well, which lies 1,500 metres (5,000ft) down on the ocean floor, about 50 miles off Louisiana.

Doug Suttles, BP's chief operating officer, insisted the operation was going to plan but admitted: "What we do know is that we have not yet stopped the flow."

He said BP engineers would soon use additional materials to try to plug the well, suggesting the heavy mud would not work on its own.

Hopes for a quick fix were further dampened today, when the company's chief executive said it would not be clear for another 48 hours whether the operation had succeeded.

Tony Hayward told CBS television's Early Show that he was 60-70% confident of success.

In a statement released today, BP said the operation could extend for another 48 hours.

It also revealed the disaster had so far cost the company $930m (640m).

Barack Obama is due to visit the Gulf later today as he continues to rebut criticism that his administration has not done enough to tackle the disaster.

Engineers began pumping thick drilling mud under pressure into the ocean floor yesterday.

The operation continues today and could extend for another 24 to 48 hours.

The spill, off the coast of Louisiana, threatens an environmental disaster and the effects will be "felt for generations to come," according to Regan Nelson, senior oceans advocate at the Natural Resources Defence Council in Washington.

He described it as an "unprecedented disaster".

May 28 2010

A delicate "top kill" operation by BP appears to have halted the gush of oil and gas from its ruptured Macondo well in the Gulf of Mexico, although experts warned that the underwater leak was still far from being permanently fixed.

Engineers were due last night to begin a second round of pumping thick drilling mud at high speed into the ocean floor, which lies 5,000 feet (1,500 metres) under the surface of the sea about 50 miles off Louisiana.

Doug Suttles, BP's chief operating officer, insisted that the operation was going to plan, but admitted: "What we do know is that we have not yet stopped the flow."

Figure 1: A demonstration system for Interactive Personalized Summarization when compression rate ϕ is specified (e.g. 5%). For convenience of browsing, we number the selected sentences (see in part 3). Extracted semantic units, such as "drilling mud", are in bold and underlined format (see in part 1). When the user clicks a sentence (part 4), the clicked sentence **ID** is kept in the *click record* (part 2). Mis-clicked records revocation can be operated by clicking the deletion icon "X" (see in part 3). Once a sentence is clicked, user can track the sentence into the popup source document to examine the contexts. The selected sentences are highlighted in the source documents (see in part 5).

Table 2: News sources of 4 datasets

News Sources	Nation	News Sources	Nation
BBC	UK	Fox News	US
Xinhua	China	MSNBC	US
CNN	US	Guardian	UK
ABC	US	New York Times	US
Reuters	UK	Washington Post	US

Table 3: Detailed basic information of 4 datasets.

News Subjects	#size	#docs	#RS	Avg.L
1.Influenza A	115026	2557	5	83
2.BP Oil Spill	63021	1468	6	76
3.Haiti Earthquake	12073	247	2	32
4.Jackson Death	37819	925	3	64

#size: total sentence counts; #RS: the number of reference summaries; Avg.L: average length of reference summary measured in sentences.

and traditional utility. User utility is obtained from interaction. The system keeps the clicked sentence records and calculates the user feedback by Equation (3) during every session. Consider sometimes users click into the summary due to confusion or mis-operations, but not their real interests. The system supports click records revocation. More details of the user interface is demonstrated in Figure 1.

4.3 Evaluation Metrics

We include both subjective evaluation from 3 evaluators based on their personalized interests and preference, and the objective evaluation based on the widely used ROUGE metrics (Lin and Hovy, 2003).

Evaluator Judgments

Evaluators are requested to express an opinion over all summaries based on the sentences which they deem to be important for the news. In general a summary can be rated in a 5-point scale, where "1" for "terrible", "2" for "bad", "3" for "normal", "4" for "good" and "5" for "excellent". Evaluators are allowed to judge at any scores between 1 and 5, e.g. a score of "3.3" is adopted when the evaluator feels difficult to decide whether "3" or "4" is more appropriate but with preference towards “3”.

ROUGE Evaluation

The DUC usually officially employs ROUGE measures for summarization evaluation, which measures summarization quality by counting overlapping units such as the N-gram, word sequences, and word pairs between the candidate summary and the reference summary. We use ROUGE-N as follows:

$$\text{ROUGE-N} = \frac{\sum\limits_{S\in\{\text{RefSum}\}}\ \sum\limits_{\text{N-gram}\in S} \text{Count}_{\text{match}}(\text{N-gram})}{\sum\limits_{S\in\{\text{RefSum}\}}\ \sum\limits_{\text{N-gram}\in S} \text{Count (N-gram)}}$$

where N stands for the length of the N-gram and N-gram$\in$RefSum denotes the N-grams in the reference summaries while N-gram$\in$CandSum denotes the N-grams in the candidate summaries. $\text{Count}_{\text{match}}$(N-gram) is the maximum number of N-gram in the candidate summary and in the set of reference summaries. $\text{Count}_{(\text{N-gram})}$ is the number of N-grams in the reference summaries or candidate summary.

According to (Lin and Hovy, 2003), among all sub-metrics in ROUGE, ROUGE-N (N=1, 2) is relatively simple and works well. In this paper, we evaluate our experiments using all methods provided by the ROUGE package (version 1.55) and only report ROUGE-1, since the conclusions drawn from different methods are quite similar. Intuitively, the higher the ROUGE scores, the similar two summaries are.

4.4 Algorithms for Comparison

We implement the following widely used multi-document summarization algorithms as the baseline systems, which are all designed for traditional summarization without user interaction. For fairness we conduct the same preprocessing for all algorithms.

Random: The method selects sentences randomly for each document collection.

Centroid: The method applies MEAD algorithm (Radev et al., 2004) to extract sentences according to the following parameters: centroid value, positional value, and first-sentence overlap.

GMDS: The Graph-based MDS proposed by (Wan and Yang, 2008) first constructs a sentence connectivity graph based on cosine similarity and then selects important sentences based on the concept of eigenvector centrality.

IPS$_{\text{ini}}$: The initial generated summary from IPS merely models *coverage* and *diversity* utility, which is similar to the previous work described in (Allan et al., 2001) with different goals and frameworks.

IPS: Our proposed algorithms with personalization component to capture *interest* by user feedbacks. IPS generates summaries via iterative sentence substitutions within user interactive sessions.

RefSum: As we have used multiple reference summaries from websites, we not only provide ROUGE evaluations of the competing systems but also of the reference summaries against each other, which provides a good indicator of not only the upper bound ROUGE score that any system could achieve, but also human inconsistency among reference summaries, indicating personalization.

4.5 Overall Performance Comparison

We take the average ROUGE-1 performance and human ratings on all sets. The overall results are shown in Figure 2 and details are listed in Tables 4∼6.

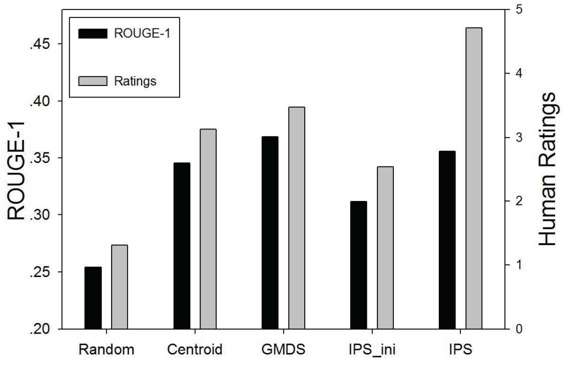


Figure 2: Overall performance on 6 datasets.

From the results, we have following observations:

• Random has the worst performance as expected, both in ROUGE-1 scores and human judgements.

• The ROUGE-1 and human ratings of Centroid and GMDS are better than those of Random. This is mainly because the Centroid based algorithm takes into account positional value and first-sentence overlap, which facilitates main aspects summarization and PageRank-based GMDS ranks the sentence using eigenvector centrality which implicitly accounts for information subsumption among all sentences.

• In general, the GMDS system slightly outperforms Centroid system in ROUGE-1, but the human judgements of GMDS and Centroid are of no significant difference. This is probably due to the difficulty

Table 4: Overall performance comparison on *Influenza A*. ROI* category: Science.

Systems	R-1	95%-conf.	H-1	H-2	H-3
RefSum	0.491	0.44958	3.5	3.0	3.9
Random	0.257	0.75694	1.2	1.0	1.0
Centroid	0.331	0.45073	2.5	3.0	3.5
GMDS	**0.364**	0.33269	3.0	2.7	3.5
IPS$_{\text{ini}}$	0.302	0.21213	2.0	2.5	2.5
IPS	0.337	0.46757	**4.8**	**4.5**	**4.5**

Table 5: Overall performance comparison on *BP Oil Leak*. ROI category: Accidents.

Systems	R-1	95%-conf.	H-1	H-2	H-3
RefSum	0.517	0.48618	4.0	3.3	3.9
Random	0.262	0.64406	1.5	1.0	1.5
Centroid	0.369	0.34743	3.2	3.0	3.5
GMDS	**0.389**	0.43877	3.5	3.0	3.9
IPS$_{\text{ini}}$	0.327	0.53722	3.0	2.5	3.0
IPS	0.372	0.35681	**4.8**	**4.5**	**4.5**

Table 6: Overall performance comparison on *Haiti Earthquake*. ROI category: Disasters.

Systems	R-1	95%-conf.	H-1	H-2	H-3
RefSum	0.528	0.30450	3.8	4.0	4.0
Random	0.266	0.75694	1.5	1.5	1.8
Centroid	0.362	0.43045	3.6	3.0	4.0
GMDS	0.380	0.33694	3.9	3.5	4.0
IPS$_{\text{ini}}$	0.331	0.34120	2.8	2.5	3.0
IPS	**0.391**	0.40069	**5.0**	**4.7**	**5.0**

Table 7: Overall performance comparison on *Michael Jackson Death*. ROI category: Legal Cases.

Systems	R-1	95%-conf.	H-1	H-2	H-3
RefSum	0.482	0.47052	3.5	3.5	4.0
Random	0.232	0.52426	1.2	1.0	1.5
Centroid	0.320	0.21045	3.0	2.5	2.7
GMDS	**0.341**	0.30070	3.5	3.3	3.9
IPS$_{\text{ini}}$	0.287	0.48526	2.5	2.0	2.2
IPS	0.324	0.36897	**5.0**	**4.5**	**4.8**

*ROI: news categorization defined by Linguistic Data Consortium. Available at http://www.ldc.upenn.edu/projects/tdt4/annotation

Table 8: Ratings consistency between evaluators: mean ± standard deviation over the 4 datasets.

RefSum	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.35±0.09	0.30±0.33
Evaluator 2			0.50±0.14

Random	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.23±0.04	0.20±0.02
Evaluator 2			0.33±0.06

Centroid	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.45±0.03	0.50±0.12
Evaluator 2			0.55±0.11

GMDS	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.35±0.02	0.35±0.03
Evaluator 2			0.70±0.03

IPS$_{\text{ini}}$	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.45±0.01	0.25±0.04
Evaluator 2			0.30±0.06

IPS	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.35±0.01	0.18±0.02
Evaluator 2			0.28±0.04

of human judgements on comparable summaries.

• The results of ROUGE-1 and ratings for IPS$_{\text{ini}}$ are better than Random but worse than Centroid and GMDS. The reason in this case may be that IPS$_{\text{ini}}$ does not capture sufficient attributes: coverage and diversity are merely fundamental requirements.

• Traditional summarization considers sentence selection based on corpus only, and hence neglects user interests. Many sentences are extracted due to arbitrary assumption of reader preference, which results in a low user satisfaction. Human judgements under our proposed IPS framework greatly outperform baselines, indicating that the appropriate use of human interests for summarization are beneficial.

The ROUGE-1 performance for IPS is not as ideal as that of GMDS. This situation may result from the divergence between user interests and general information provided by mass media propaganda, which again motivates the need for personalization.

Although the high disparities between different human evaluators have been observed in (Gong and Liu, 2001), we still examine the consistency among 3 evaluators and their preferred summaries to prove the motivation of personalization in our work.

4.6 Consistency Analysis for Personalization

The low ROUGE-1 scores of RefSum indicate the inconsistency among reference summaries. We conduct personalization analysis from two perspectives: (1) human rating consistency and (2) content consistency among human supervised summaries.

We calculate the mean and variance of rating variations among evaluator judgements, listed in Table

Table 9: Content consistency among evaluators supervised summaries.

	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1		0.273	0.398
Evaluator 2	0.289		0.257
Evaluator 3	0.407	0.235	
RefSum	0.365	0.302	0.394

8. We see that for Random the average rating variation is 0.25, for IPS is 0.27, for IPS$_{\text{ini}}$ is 0.33, for RefSum is 0.38, for GMDS is 0.47 and for Centroid is the highest, 0.50. Such phenomenon indicates for poor generated summaries, such as Random or IPS$_{\text{ini}}$, humans have consensus, but for normal summaries without personalized interests, they are likely to have disparities, surprisingly, even for RefSum. General summaries provided by mass media satisfy part of audiences, but obviously not all of them.

The high rating consistency of IPS indicates people tend to favor summaries generated according to their interests. We next examine content consistency of these summaries with high rating consistency.

As shown in Table 9, although highly scored, these human supervised summaries still have low content consistency (especially Evaluator 2). The low content consistency between RefSum and supervised summaries shows reader have individual personalization. Note that the inconsistency among evaluators is larger than that between RefSum and supervised summaries, indicating *interests* take a high proportion in evaluator supervised summaries.

4.7 Parameter Settings

δ controls coverage/diversity tradeoff. We tune δ on IPS$_{\text{ini}}$ and apply the optimal δ directly in IPS. According to the statistics in (Yan et al., 2010), the semantic coherent context is about 7 sentences. Therefore, we empirically choose k=3 for the examined context window. The number of topics is set at n=50. We assign an equal weight ($\gamma = 1$) to semantic units and examined contexts according to analogical research of summarization from implicit feedbacks via clickthrough data (Sun et al., 2005).

λ is the key parameter in IPS approach, controlling the weight of user utility during the process of interactive personalized summarization.

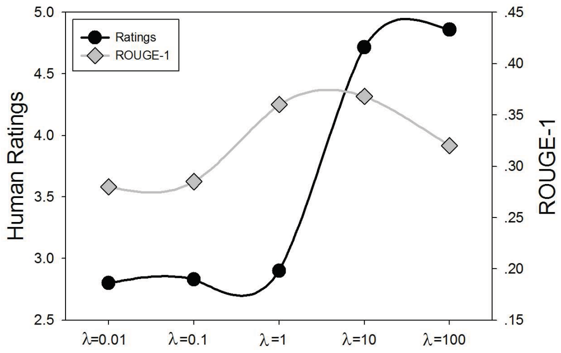


Figure 3: λ v.s. human ratings and ROUGE scores.

Through Figure 3, we see that when λ is small ($\lambda \in [0.01, 0.1]$), both human judgements and ROUGE evaluation scores have little difference. When $\lambda \in [0.1, 1]$, ROUGE scores increase significantly but human satisfaction shows little response. $\lambda \in [1, 10]$ brings large user utility enhancement because user may find what they are interested in but ROUGE scores start to decay. When $\lambda \in [10, 100]$, ROUGE scores drop much because the emphasized user interests may guide the generated summaries divergent away from the original corpus.

In Figure 4 we examine how λ attracts user clicks and regeneration counts until satisfaction. As the result indicates, both counts increase as λ increases. When λ is small (from 0.01 to 0.1), readers find no more interesting aspects through clicks and regenerations and stop due to the bad user experience. As λ increases, the system mines more relevant sentences according to personalized interests and hence attracts user clicks and intention to regenerate.

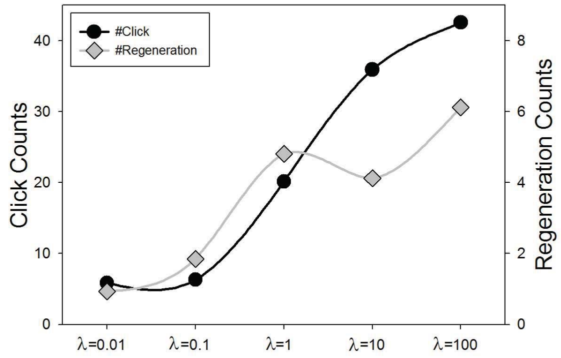


Figure 4: λ v.s. click counts and regeneration counts.

5 Conclusion

We present an important and novel summarization problem, Interactive Personalized Summarization (IPS), which generates summaries based on human−system interaction for "interests" and personalization. We formally formulate IPS as a combination of user utility and traditional summary utility, such as coverage and diversity. We implement a system under such framework for experiments on real web datasets to compare all approaches. Through our experiments we notice that user personalization of interests plays an important role in summary generation, which largely increase human ratings due to user satisfaction. Besides, our experiments indicate the inconsistency between user preferred summaries and reference summaries measured by ROUGE, and hence prove the effectiveness of personalization.

Acknowledgments

This work was partially supported by HGJ 2010 Grant 2011ZX01042-001-001 and NSFC with Grant No.61073082, 60933004. Rui Yan was supported by the MediaTek Fellowship.

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of the 24th annual international SIGIR'01*, pages 10–18.

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

G. Erkan and D.R. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP'04*, volume 4.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of SIGIR'99*, pages 121–128.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th international ACM SIGIR conference*, SIGIR '01, pages 19–25.

Q. Guo and E. Agichtein. 2010. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceeding of the 33rd international ACM SIGIR conference*, SIGIR'10, pages 130–137.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR'04*, pages 297–304.

Anton Leuski, Chin-Yew Lin, and Eduard Hovy. 2003. ineats: interactive multi-document summarization. In *Proceedings of ACL'03*, pages 125–128.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. 2009. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of WWW'09*, pages 71–80.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization: a prototype system and its evaluation. In *Proceedings of ACL'02*, pages 457–464.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL'03*, pages 71–78.

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *EMNLP'09*, pages 306–314.

R. Mihalcea and P. Tarau. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*, volume 5.

D.R. Radev, H. Jing, and M. Sty. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In *Proceedings of SIGIR'05*, pages 194–201.

Stephen Wan and Cécile Paris. 2008. In-browser summarisation: generating elaborative summaries biased towards the reading context. In *ACL-HLT'08*, pages 129–132.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of SIGIR'08*, pages 299–306.

X. Wan, J. Yang, and J. Xiao. 2007a. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI*, volume 7, pages 2903–2908.

X. Wan, J. Yang, and J. Xiao. 2007b. Single document summarization with document expansion. In *Proceedings of the 22nd AAAI'07*, pages 931–936.

Rui Yan, Yu Li, Yan Zhang, and Xiaoming Li. 2010. Event recognition from news webpages through latent ingredients extraction. In *AIRS'10*, pages 490–501.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th annual international ACM SIGIR'11*.

Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achanauparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical Keyphrase Extraction from Twitter. In *Proceedings of ACL-HLT'11*.