Social Context Summarization

Zi Yang∗, Keke Cai†, Jie Tang∗, Li Zhang∗, Zhong Su† and Juanzi Li∗

∗Department of Computer Science and Technology, Tsinghua University, China
† IBM, China Research Lab

{yangzi, tangjie, ljz}@keg.cs.tsinghua.edu.cn, {caikeke, lizhang, suzhong}@cn.ibm.com

ABSTRACT

We study a novel problem of social context summarization for Web documents. Traditional summarization research has focused on extracting informative sentences from standard documents. With the rapid growth of online social networks, abundant user generated content (e.g., comments) associated with the standard documents is available. Which parts in a document are social users really caring about? How can we generate summaries for standard documents by considering both the informativeness of sentences and interests of social users? This paper explores such an approach by modeling Web documents and social contexts into a unified framework. We propose a dual wing factor graph (DWFG) model, which utilizes the mutual reinforcement between Web documents and their associated social contexts to generate summaries. An efficient algorithm is designed to learn the proposed factor graph model. Experimental results on a Twitter data set validate the effectiveness of the proposed model. By leveraging the social context information, our approach obtains significant improvement (averagely +5.0%–17.3%) over several alternative methods (CRF, SVM, LR, PR, and DocLead) on the performance of summarization.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.2.8 [Database Management]: Data Mining; J.4 [Computer Applications]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation

Keywords

Document summarization, Social context, Factor graph, Twitter

1. INTRODUCTION

Web document summarization has been widely studied for many years. Existing methods mainly use statistical or linguistic information such as term distribution, sentence position, and topics to extract the most informative sentences from standard (Web) documents. However, these methods only consider the document’s content information, but ignore how users (readers) think about the document. With the rapid growth of online social networks, users can freely express what they are thinking about any Web document. For example, many news websites allow the users to directly add comments to each news document. On Twitter1, many users post the URL address of a news document onto their microblogs, followed by some personal comments. The comments imply the importance of different sentences and can be used to help improve the quality of document summarization. More importantly, the users’ comments essentially reflect which part of the document that they are interested in.

In this work, we present a novel problem of social context summarization. The question we intend to answer is: how to generate a summary for Web documents by considering both the informativeness of sentences and interests of social users? The concept of “context” for summarization has been previously studied and various approaches have been proposed based on different kinds of context, such as hyperlinks [1, 9], click-through data [29], comments [13, 19], or opinionated text [26, 11, 15]. Most of these methods directly integrate the context information into the target Web page to help estimate the informativeness of sentences. However, in this way, the context information is only considered as textual information. Many important information has been ignored. For example, if a user is an opinion leader, his comments should be more important than others. From a comment’s perspective, if a comment has been forwarded or replied by many other users, the comment should be more important than others. One goal of this work is to consider the social influence and the information propagation for document summarization. The problem is referred to as social context summarization. The problem is clearly different from existing research and poses a set of unique challenges:

• First, social context is becoming more and more complicated. There are users, user generated contents, social networks, and implicit networks (such as the forward/reply network). How to formally define the social context is a non-trivial problem.

• Second, in the social environment, the quality of summaries strongly depends on the social context information (such as the social influence between users). How to formalize the problem in a principled framework is a challenging problem.

• Third, the social context contains inevitable noise. How to qualitatively analyze the problem and quantitatively validate

1http://www.twitter.com, a microblogging system.
In this paper, we try to systematically investigate the problem of social context summarization for Web documents. We formulate the problem of social context summarization and propose a dual-wing factor graph (DWFG) model. The DWFG model incorporates the summarization task and the social network analysis into a unified framework. In this way, the two tasks can be mutually reinforced. We employ the Microblogging as an example to quantitatively study the social context summarization problem. In particular, we crawl a data set from Twitter. The user generated content is the tweet posted by the user. The retweeting (forwarding) and replying relationships between tweets form an implicit information network, and the following relationships between users form the user network. Some tweets have the links pointing to standard Web documents (such as news documents). The problem then becomes how to leverage both the information network and the user network to generate high-quality summaries for standard Web documents.

The overview of the proposed method is a supervised framework (as shown in Figure 1). In training, we estimate the importance of defined features and strength of dependencies for identifying key sentences and microblogs in the social contexts. In test, given a new Web document with its social context, we perform collective inference for the importance of each sentence and microblog and select a subset of sentences as the summary according to the trained models. We validate the proposed approach on the Twitter data set. The experiment results show that by leveraging the social context information, our approach can significantly improve (on average +10.8%) the performance of summarization. We also compare with a set of alternative methods (i.e., CRF, SVM, LR, PR, and DocLead), and our approach clearly outperforms (averagely +5.0%-17.3%) the baseline methods.

The paper is organized as follows: in Section 2, we introduce some notations and formally define the problem. In Section 3, we propose the factor graph model to address the problem, and in Section 4, we conduct the experiments on the Twitter dataset. Finally, in Section 5 and 6, we summarize related works and conclude.

2. PROBLEM DEFINITION

In this section, we introduce some notations and representations necessary for the problem, and then define the problem of social context summarization.

Definition 1 (Social context). Given a Web document $d$, its social context $C_d$ is defined as $(M_d, U_d)$, where $M_d$ is a set of comments on $d$ written by users $U_d$ in a social network.

In the context of Web 2.0, Web documents, e.g., news or blogs, are freely discussed and commented by users. These comments again spread (e.g., by forwarding between friends) in the social network. The users’ activities implicitly reflect the importance of different parts (e.g., sentences) in the document from the user’s perspective. We believe that the social context, Thus, integrating the document content information and the social context information can disclose a more thorough view of the document. In this paper, we employ Twitter as the basis for our study. Specifically, given a Web document $d$ and its associated social context (tweets $M_d$ containing the URL address of document $d$ and users $U_d$ who posted those tweets), we give the following definition of social context augmented network.

Definition 2 (Social Context Augmented Network, SCAN). Social Context Augmented Network $G_d = (S_d, C_d, E_d)$ is defined as a network that is built upon the sentence set $S_d$ of document $d$ and its social context $C_d$, where the edge set $E_d$ contains three types of edges: $E_d^1$, $E_d^2$, and $E_d^3$. $E_d^1$ represents the relationships between document sentences, $E_d^1 = \{(s_i, s_j) | s_i, s_j \in S_d\}$ represents the relationships between messages, and $E_d^3 = \{(u_i, u_j) | u_i, u_j \in U_d\}$ represents the relationships between users.

Compared with traditional contexts that are defined based on textual information, social context need model various dynamic social relationships, such as the follower-followee relationships between users, retweeting relationships and replying relationships between tweets. An example of SCAN is shown in Figure 2(a).

In this figure, the upper layer includes two documents $d_1$ and $d_2$ and $d_1$ contains four sentences $s_1$, $s_2$, $s_3$, and $s_4$. The two documents are respectively associated with two sets of messages $M_1 = \{m_1, m_2, m_3, m_4\}$ and $M_2 = \{m_5, m_6\}$ in the middle layer. The lower layer refers to the user layer consisting of users $u_1$, $u_2$, and $u_3$, who are also associated with the messages $M_1$ and $M_2$. Besides external relationships between the objects across different layers, SCAN also describes internal relationships between objects within the same layer (as shown in Figure 2(a)).

Given this, we can formally define our problem of Social Context Summarization.

Definition 3 (Social Context Summarization). Given a social context augmented network $G_d$, the goal of social context summarization is to generate a summary which consists of two pieces of information: the most important sentences $S_d^3 \subseteq S_d$ and the most representative messages $M_d \subseteq M_d$.

The problem of social context summarization contains two subproblems: Key Sentence Extraction and Tweet Summarization. In the former problem, we aim to identify the most important sentences from document $d$'s content, while in the latter subproblem, we intend to find the most representative tweets from the social context $C_d$ of document $d$. Social context $C_d$ contains rich information about the document $d$, which is helpful for the Key Sentence Extraction problem, while the important sentences in a document can equally help Tweet Summarization in the social context. The mutual reinforcement between the two subproblems can facilitate generating a high-quality summary. Moreover, social context summarization could also answer a number of related questions, e.g., who are the most experienced users of a specific topic or a fact mentioned in a document.
3. PROBLEM SOLVING

In this section, we propose a dual wing factor graph (DWFG) model, which formulates the social context summarization problem in a unified learning framework. The DWFG model simultaneously incorporates all resources in social context to generate high-quality summaries for Web documents.

3.1 Basic Idea

In our Twitter data set, each Web document is associated with a social context. To generate summaries for Web documents, a straightforward method is to define a set of features to characterize the importance of each sentence, and then use a classification model to identify which sentences should be included into the summary [16, 25, 36]. To further consider the correlation between sentences, we can consider a sequential labeling approach such as conditional random field. Such a method has been studied by [8, 28]. Both of them consider the sentence local features and similarities (correlations) between sentences, and model the sentence extraction task with a linear-chain conditional random field. An example of the graphical representation is shown in Figure 2(b). The method only considers the correlation between sentences (the document layer in Figure 2(a)), but ignores the social context information resided in the microblog and user layers.

To model the tweet network, we design another similar graphical model with structures reflecting the information propagation. Figure 2(c) presents an example. Each gray circle indicates a tweet, the arrow represents the replying/retweeting relationship between two tweets. Based on such a formulation, we can define local features (content-based features) for each tweet, as well as edge features for each replying/retweeting relationship. By learning such a graphical model, we can classify which tweets are important (or informative). Obviously, this model only considers the information from the tweet side and does not consider the Web documents. An ideal way is to incorporate the two tasks together so that they can reinforce each other.

Based on these considerations, we propose a novel dual wing factor graph (DWFG) model. The graphical representation is shown in Figure 2(d). In the DWFG model, the upper layer is used to model the key sentence extraction task and the bottom layer is designed to model the tweet summarization problem. In the middle layer, we design a set of correlation factor functions $h$ to bridge the two tasks. By carefully designing the correlation factor function $h$, we can elegantly combine the two tasks of key sentence extraction and tweet summarization into a unified framework. In the rest of this section, we will explain in details how we design and learn the dual wing factor graph model.

3.2 Modeling Summarization via Dual Wing Factor Graphs

We model the social context summarization problem in the dual wing factor graph (DWFG) model. Each sentence $s_i \in S_d$ or tweet $m_j \in M_d$ is associated with a binary value $y_i$ indicating the importance of the sentence or tweet (1 representing important, and 0 representing unimportant).

We first collect a set of labeled SCANS (training set) $T = \{G_d\}_{d=1}^D$, i.e., each sentence $s_i \in S_d$ and tweet $m_j \in M_d$ in each social context $C_d$ are associated with known binary labels $y_i$ and $y_j$, moreover, we also collect the test set $S$ of unlabeled instances, which consists of all the sentences and tweets not yet judged. Our goal is then to learn a DWFG model from the training set and apply it to predict which sentences and tweets are important in the test set $S$, i.e., to infer the value (label) of $y$, and then generate a summary for the social context.

We define three types of factor functions associated with individual instances or instance groups: local attribute factor, intra-domain dependency factor, and inter-domain dependency factor.

Local attribute factor. The probability that a sentence or tweet is important could be estimated by some local attributes (represented as $x$), which refer to features that are inherent to the sentence or tweet itself. In general, we define similar features for sentences and tweets. The features include the average TF-IDF score over words and the log likelihood generated by the context, the position of the sentence in the document, the author’s authoritativeness. Details of the defined local features for sentences and tweets are given in Section 4.

To estimate the significance of each feature, we introduce a weight variable $\lambda_c$ for each feature $c$, and we define a local at-
tribute factor \( f_{i,c} \) for the feature \( c \) of each sentence \( s_i \) or tweet \( m_t \). Formally, a factor could be defined as the local entropy:

\[
f_{i,c}(\lambda_c, y_i) = \exp(\lambda_c x_{i,c} y_i)
\]

where \( x_{i,c} \) is the value of the \( c \)-th feature extracted from sentence \( s_i \) or tweet \( m_t \).

**Intra-domain dependency factor.** As described in Section 3.1, we introduce factors that are capable of handling multiple instances in either sentence level or tweet level, to characterize the dependencies among sentences and tweets respectively. Intra-domain interaction may promote some sentences to become more important while inhibit others from becoming important. We associate each type of interaction with a weight \( \mu_c \) indicating the confidence of the corresponding interaction. The interaction has a positive influence only if the weight \( \mu_c \) is greater than 0. We introduce factor \( g_{i,j,c} \) to capture the dependency among sentence pair \( s_i \) and \( s_j \) or tweet pair \( m_t \) and \( m_{t_j} \).

\[
g_{i,j,c}(\mu_c, y_i, y_j) = \begin{cases} \exp \mu_c & \text{if some condition holds} \\ 1 & \text{otherwise} \end{cases}
\]

A document can be regarded as a sequence of sentences, and thus key sentence extraction could be viewed as a sequence labeling process [28], i.e., the judgment on a certain sentence is affected by the nearby sentences to avoid both sentences of high similarity are chosen simultaneously. Hence, the dependency conditions in Eq. 2 for a sentence pair \( s_i \) and \( s_j \) can be formalized as follow: the factor takes value \( \exp \mu_c \) if \( y_i \neq 1 \) or \( y_j \neq 1 \). To avoid high computational complexity, we only constrain consecutive and similar sentences, i.e., establish sentence relation for sentence \( s_i \) and \( s_{i+1} \) whose mutual similarity (e.g., cosine similarity) exceeds the threshold \( \theta_q \).

Moreover, we consider the two interactions among tweets: replying and retweeting. If tweet \( m_t \) replies or retweets tweet \( m_{t_j} \), then \( m_t \) successfully excites and attracts attentions from others, and it is reasonable that \( m_t \) is more important than its succeeding tweets in the thread. Formally, for such a tweet pair \( m_t \) and \( m_{t_j} \), the factor takes value \( \exp \mu_c \) if \( y_i \leq y_j \).

**Inter-domain dependency factor.** By leveraging knowledge from both domains, the inter-domain relationships may benefit to the identification of social context summarization. We introduce a set of factors defined on variables across domains, which are able to coordinate the labels of sentences and tweets simultaneously. Specifically, if tweet \( m_t \) is considered as a representative tweet, i.e., \( y_j = 1 \), then a sentence \( s_i \) highly similar to \( m_t \) (with similarity more than a threshold \( \theta_h \)) should be biased towards the same label, i.e., \( y_i = 1 \). Formally, for each sentence-tweet pair \((s_i, m_j)\) of high similarity, we define

\[
h_{i,j}(\nu, y_i, y_j) = \begin{cases} \exp \nu & \text{if } y_i = y_j \\ 1 & \text{otherwise} \end{cases}
\]

where \( \nu \) is the weight variable that represents the significance of inter-domain dependency factor.

**Objective function.** Finally, the objective function can be defined as the normalized product of Eqs. 1 - 3 for all the instances. We denote \( Z \) as the normalization factor, which sums up the conditional likelihood \( P(Y | X, \Theta) \) over all the possible labels of all the instances, where \( Y \) contains all the undetermined labels for sentences and tweets, i.e., \( Y = \{y_i\}_i \), and \( \Theta \) is the collection of weights, i.e., \( \Theta = \{\lambda_c\}_c \cup \{\mu_c\}_c \cup \{\nu\}_\nu \).

We first estimate the parameters \( \Theta \) with a maximum likelihood procedure on the training instances, e.g.,

\[
\max_{\Theta} \frac{1}{Z} \prod_{i,j \in T} \prod_{c \in C} f_{i,c}(\lambda_c, y_i) \cdot g_{i,j,c}(\mu_c, y_i, y_j) \cdot h_{i,j}(\nu, y_i, y_j)
\]

We use L-BFGS, a quasi-Newton method for solving the non-linear optimization problem (i.e., Eq. 4). To avoid overfitting, we add a penalty term \(-\frac{1}{2}||\Theta||^2 / \sigma^2\), a spherical Gaussian prior, into the objective function, which is a regularization method commonly used in maximum entropy and conditional random fields [6, 27, 28].

Calculating the marginal distribution for each factor (in deriving the log-gradient of the objective function) requires a loopy sum-product inference algorithm. With the learned parameter \( \Theta \), we may summarize an unlabeled social context for a Web document in the test set by extracting important sentences, which are also identified by a similar max-sum inference procedure. The inference algorithm is introduced in the next section.

**Connection with existing models.** We note that the proposed DWFG model can also be viewed as a model generalized from existing models. In Eq. 4, if parameter \( \nu \) is fixed as \( 0 \), i.e., all factors \( \{h_{i,j}\}_{ij} \) take constant values of 1, and factors \( \{f_{i,c}\}_{i,c} \) and \( \{g_{i,j,c}\}_{i,j,c} \) are only defined for sentences, then the simplified model only incorporates sentence local factors and sentence relation factors, and DWFG model is degenerated to a special case: the summarization approach based on linear-chain CRF [28]. Moreover, if all parameters \( \{\mu_c\}_c \) are also set as \( 0 \), i.e., only the local factors \( \{i,c\}_i,c \) are non-trivial, then DWFG is turned into the logistic regression classifier [25].

### 3.3 Inference Algorithm

Since the graphical model DWFG proposed for summarization (cf. Figure 2(d)) contains cycles, we cannot directly employ a forward-backward algorithm like in [28] for exactly inferring the optimal labeling for a test instance. We then propose an approximate inference approach based on the loopy sum-product or max-sum algorithm.

To achieve an approximate inference for predicting labels, the algorithm contains multiple iterations for updating the beliefs, and each iteration is comprised of two phases. Here, we denote the update variables for delivering beliefs between variables and factors by \( p_{i,c} \) and \( q_{i,j} \). \( \{p_{i,c}\}_{ic} \) represent the messages propagating from variable (e.g., \( y_i \)) to factor (e.g., \( g_{i,j,c} \) or \( h_{i,j} \)), and \( \{q_{i,j}\}_{ij} \) represents the messages factor to variable respectively. The messages can be formulated as follows:

\[
p_{i,j} = r_i + \sum_{k \in N(i) \setminus \{j\}} q_{ik}
\]

\[
q_{i,j} = \max\{t_{i,j}(1,1) + p_{ij}, t_{i,j}(1,0)\} - \max\{t_{i,j}(0,1) + p_{ij}, t_{i,j}(0,0)\}
\]

where \( r_i \) corresponds to the logarithmic value of the local factor, i.e., \( r_i = \sum_{c \in C} (\log f_{i,c}(\lambda_c, y_i = 1) - \log f_{i,c}(\lambda_c, y_i = 0)) \). Analogously, \( t_{i,j}(y_i, y_j) \) is the logarithmic value of the dependency factor, i.e., \( t_{i,j}(y_i, y_j) = \log g_{i,j,c}(\mu_c, y_i, y_j) \) or \( h_{i,j}(\nu, y_i, y_j) \). Specific to a particular dependency factors, \( f_{i,c}, g_{i,j,c}, \) or \( h_{i,j} \) (Eq. 1 to 3), the message \( q_{i,j} \) has a more succinct expression, e.g., the sentence dependency factor \( q_{i,j} = \max\{p_{ij}, -\mu_c\} - \max\{p_{ij}, 0\} \).

We can obtain the label for each sentence \( s_i \) and tweet \( m_t \) using the variables calculated in the two phases for the last iteration as follows:

\[
y_i = \begin{cases} 1 & \text{if } p_{ij} + q_{i,j} > 0 \text{ for some } j \\ 0 & \text{otherwise} \end{cases}
\]
Algorithm 1: Social context summarization with DWFG

\textbf{input}: A document \( d \) with social context \( C_d \) and SCAN \( G_d \) of \( d \), weight variables \( \Theta \), and number of iterations \( I \)

\textbf{output}: A summary for social context \( C_d \); important sentences \( S_d^t \) and messages \( M_d^t \)

// initialization
1 initialize variables \( \{r_i\} \leftarrow 0 \);
// update message values
2 for \( i \leftarrow 1 \) to \( I \) do
3 update variables \( p_{ij} \) according to Eq. 5;
4 update variables \( q_{ij} \) according to Eq. 6;
// output result
5 foreach \( s_i \in S_d \) and \( m_i \in M_d \) do
6 calculate \( y_i \) according to Eq. 7;
7 \( S_d^t \leftarrow \{s_i \in S_d | y_i = 1\} \);
8 \( M_d^t \leftarrow \{m_i \in M_d | y_i = 1\} \);

The learning algorithm is depicted in Algorithm 1. Initially, we calculate all local variables \( \{r_i\} \), and initialize all update variables \( \{q_{ij}\}_{i,j} \) as 0 (Line 1). Next, we compute new values for all the update variables \( \{p_{ij}\}_{i,j} \) according to Eq. 5. Then we estimate the new values for all \( \{q_{ij}\}_{i,j} \) according to Eq. 6. We continue to update the variables for a number of iterations until some termination condition is satisfied. Finally, the summary of the social context is generated according the update variables (Line 5 - 10).

\textbf{Complexity analysis}. If we denote the number of iterations for the inference algorithm as \( I \), then the computational complexity of the algorithm is proportional to \( I \cdot (|E^t| + |E^m| + |E^r|) \), where \( |E^t|, |E^m|, |E^r| \) correspond to the number of sentence relationships, tweet relationships, and inter-domain relationships respectively. They can be varied from zero to many when we tune the thresholds \( \theta_y \) and \( \theta_h \), which is further discussed in Section 3.2. In fact, the inference algorithm can be easily parallelized or distributed onto clusters to handle large-scale dataset, and the design of distributed algorithm will be reported elsewhere.

4. EXPERIMENT

In this section, we evaluate the proposed summarization method DWFG with manually labeled documents. We firstly introduce the data set, baseline methods that do not incorporate the relationship between the Web document domain and tweet thread domain, the evaluation metrics, and then we give the detailed discussion of the experiment results with the comparison of other approaches. More supplied materials of this work can be found at http://arnetminer.org/socialcontext/.

4.1 Settings and Observations

4.1.1 Data Preparation

Since there is rarely previous work study the summarization task from social perspective, to the best of our knowledge no existing benchmark dataset can be utilized for our experiments. We collected data from the most popular microblogging website, Twitter.

From 4,874,389 Twitter users, we collected 404,544,462 tweets within a period from January 1st 2010 to July 17th 2010, and then recognized all the tweets accompanied with explicit URLs (containing “http://” or “https://”). Since users might use different URL shortening services, such as tinyurl.com, bit.ly, etc., we simply implemented a general shorter URL decoder based on a HTTP client to obtain the decoded URLs. Finally, the top 200,000 high frequent URLs were finally extracted and 12,964,166 tweets talking about the same URL were grouped together. In our preliminary experiments, we observed the distribution of frequencies of URLs carried by the tweets, which is plotted in Figure 3 in log-logarithmic scale.

We see that the highest frequent URL is mentioned by 114,911 tweets in our experiment data, and the distribution of frequencies of URLs follows the power law. According to the selected URLs, we crawled the associated Web pages, and then constructed two kinds of data sets (Web pages and their corresponding social tweets). The Web documents were then segmented into a set of sentences with the jTokenizer Toolkit. Our summarization algorithm was then performed on both domains.

We found that most of the Top 50 URLs correspond to advertisement pages. We therefore predefined a series of high-quality websites, such as CNN, BBC, Mashable etc., and selected a subset of URLs related to these websites for manual annotation. We note that a Web document might be referred by different URLs even if the URLs are decoded, e.g., URLs “http://news.bbc.co.uk/1/hi/england/8604663.stm”, “http://news.bbc.co.uk/2/hi/uk_news/england/8604663.stm”, and “http://news.bbc.co.uk/2/hi/uk_news/england/8604663.stm” correspond to the same Web document. We further group such Web documents according to the unique document ID indicated in the URL (e.g., 8604663). Details on the five selected domains are given in Table 1.

4.1.2 Evaluation Methods

To guarantee the low noise of the manual annotation data, we further manually validated the informativeness of all the selected Web documents by posting both the Web documents and tweets on Amazon Mechanical Turk\(^3\).

We totally issued 1145 HITs on Mechanical Turk, and for each HIT we asked at least two different workers to read both the Web documents and its corresponding tweets. All the HITs were divided into 12 batches with each assignment entitled “Key sentences and tweets extraction from news and related tweets”. We gave a detailed description on how to label the sentences and tweets, and also emphasized that the workers should “extract severa sentences from news that attract them mostly”, and “after reading the news, extract the most interesting tweets that appeal you mostly”. We required the workers to label no less than 5 tweets and 10 Web document sentences. Finally, 158 different users have participated in annotating the benchmark for social context summarization task.

The labeled sets of sentences and tweets formed the benchmark for evaluation.\(^4\)

\(^3\)http://code.google.com/p/jtokeniser/
\(^4\)http://mturk.com, an Internet marketplace to use human intelligence to solve various kinds of problems

\(^4\)We plan to gradually publish the annotated dataset for academic use of social context summarization.
In this paper, two performance metrics applied in [29] were adopted to evaluate the proposed approach DWFG. The first is Precision, Recall and F-measure. In the following section, we will report the evaluation on $F_1$ measure, which is defined as:

$$P = \frac{S_{ref} \cap S_{cand}}{S_{cand}}; \quad R = \frac{S_{ref} \cap S_{cand}}{S_{ref}}; \quad F_1 = \frac{2PR}{P + R}$$

where $S_{cand}$ and $S_{ref}$ denote the sentences contained in the candidate summary and the reference summary respectively.

Another performance metric is ROUGE [18], which measures summarization quality according to the overlap between the units, such as n-gram (referred to as ROUGE-N) etc, of machine generated summary and human generated summary. ROUGE-N is defined as follows:

$$\text{ROUGE-N} = \sum_{n \in S_{ref}} \sum_{gram_n \in S_{cand}} \frac{\text{Count}_{match}(gram_n)}{\text{Count}(gram_n)}$$

where $n$ is the length of the n-gram, $\text{Count}_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries, $\text{Count}(gram_n)$ is the number of n-grams in the reference summaries.

We employ the ROUGE evaluation methods implemented in the Dragon Toolkit Project6, and report the experimental results in terms of ROUGE-1 and ROUGE-2 with stop words filtered out. Since ROUGE is a recall-oriented metric, we keep the number of sentences extracted be equal with that of the human summary for fair comparison. Specifically, we select the sentences and tweets with the greatest positive beliefs by $P(y_i = 1|X)$ (cf. Eq. 7).

### 4.1.3 Feature Description

Many features have been designed for document summarization in prior literatures. In this paper, we only extract 11 basic and forward-forward features from both domains. Besides of some features that are widely used in traditional summarization methods, we also utilize several features extracted from users’ online social behaviors, e.g., the number of users following the tweet’s author and the PageRank score of the author. Table 2 gives the brief definition of these features applied in this paper, where some features were represented by nominal values, e.g., Feature 1 will take value 4 if the sentence was extracted from the title of the document, 3 if it was extracted from the subtitle, 2 if the sentence was located in the first paragraph of the original document, 1 if the sentence was located in the last paragraph, and 0 otherwise.

The feature values extracted from sentence domain and tweet domain are summarized in Figure 4 and 5. Since different features take values in diverse ranges, e.g., the maximum value of Feature 4 is 15, while the maximum value of Feature 6 is 1.495, we normalize the feature values by the mean value of corresponding feature.

![Figure 4: Comparison of feature values for sentence domain on five domains](image)

From Figure 4, we can see that Web documents from different domains exhibit differently. For example, articles in CNN, BBC, and ESPN have smaller values of Feature 1 but greater values of Feature 2 than MTV, which indicates that news Websites CNN, BBC, and ESPN have longer articles consisting of a greater number of shorter paragraphs. Therefore, we trained an independent model on each domain respectively to capture the distinctiveness.

### 4.1.4 Baseline Methods

We compare DWFG with six supervised baselines methods. SVM classifiers (SVM+) and logistic regression classifiers (LR+) are performed for each sentence and tweet only with its local features. Linear-chain and tree-structured CRF models (LC-/TS-CRF) are respectively trained and tested on documents and tweet threads, i.e., inter-domain relationships are considered as a supplement to the basic local features. The linear-chain CRF baseline model employed in the sentence summarization is equivalent to the method proposed in [28].

We also extend the feature list for each sentence and tweet by considering the features of related sentences or tweets extracted from both domains (denoted as SVM++, LR++). Specifically, for each sentence $s_i$ in a document, we append 11 features ($x_{i,1}, \ldots, x_{i,17}$), where each of $x_{i,7}, \ldots, x_{i,12}$ adds up the corresponding feature values of its similar sentences, and each of $x_{i,13}, \ldots, x_{i,17}$ adds up the corresponding feature values of its related tweets. Similarly, for each tweet in the thread, we append 11 features, which are the sums of feature values of its relevant sentences or tweets.

In addition, we also compare DWFG with commonly applied unsupervised summarization algorithms, i.e., the importance sentences and tweets are selected according to a metric or score. First, we randomly select sentences or tweets (Random) as the basic unsupervised method. Another baseline method for summarization is to select the sentences according to their positions in the document or paragraph (DocLead and ParaLead). Finally, we apply PageRank algorithm for summarization on the whole graph consisting of three types of relationships (PR) [24].

![Table 2: Feature list](image)

![Table 1: Description on employed domains](image)

---

6http://dragon.ischool.drexel.edu/
Figure 5: Comparison of feature values for tweet domain on five domains

Table 3: Experimental results for Web documents

<table>
<thead>
<tr>
<th>Method</th>
<th>CNN</th>
<th>BBC</th>
<th>MTV</th>
<th>ESPN</th>
<th>Mash</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.288</td>
<td>0.322</td>
<td>0.490</td>
<td>0.337</td>
<td>0.321</td>
<td>0.351</td>
</tr>
<tr>
<td>LR</td>
<td>0.284</td>
<td>0.340</td>
<td>0.351</td>
<td>0.352</td>
<td>0.297</td>
<td>0.361</td>
</tr>
<tr>
<td>LC-CRF</td>
<td>0.307</td>
<td>0.349</td>
<td>0.596</td>
<td>0.364</td>
<td>0.340</td>
<td>0.391</td>
</tr>
<tr>
<td>SVM+</td>
<td>0.283</td>
<td>0.341</td>
<td>0.476</td>
<td>0.359</td>
<td>0.324</td>
<td>0.357</td>
</tr>
<tr>
<td>LR+</td>
<td>0.277</td>
<td>0.332</td>
<td>0.482</td>
<td>0.366</td>
<td>0.305</td>
<td>0.352</td>
</tr>
<tr>
<td>F$_1$ Random</td>
<td>0.314</td>
<td>0.321</td>
<td>0.455</td>
<td>0.351</td>
<td>0.305</td>
<td>0.349</td>
</tr>
<tr>
<td>DocLead</td>
<td>0.334</td>
<td>0.356</td>
<td>0.441</td>
<td>0.317</td>
<td>0.415</td>
<td>0.373</td>
</tr>
<tr>
<td>ParaLead</td>
<td>0.298</td>
<td>0.316</td>
<td>0.508</td>
<td>0.338</td>
<td>0.323</td>
<td>0.356</td>
</tr>
<tr>
<td>PR</td>
<td>0.354</td>
<td>0.338</td>
<td>0.453</td>
<td>0.351</td>
<td>0.399</td>
<td>0.379</td>
</tr>
<tr>
<td>DWFG</td>
<td>0.341</td>
<td>0.450</td>
<td>0.642</td>
<td>0.518</td>
<td>0.330</td>
<td>0.456</td>
</tr>
</tbody>
</table>

Table 4: Experimental results for tweet thread

<table>
<thead>
<tr>
<th>Method</th>
<th>CNN</th>
<th>BBC</th>
<th>MTV</th>
<th>ESPN</th>
<th>Mash</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.531</td>
<td>0.542</td>
<td>0.640</td>
<td>0.610</td>
<td>0.379</td>
<td>0.499</td>
</tr>
<tr>
<td>LR</td>
<td>0.370</td>
<td>0.531</td>
<td>0.606</td>
<td>0.616</td>
<td>0.408</td>
<td>0.506</td>
</tr>
<tr>
<td>LC-CRF</td>
<td>0.378</td>
<td>0.547</td>
<td>0.637</td>
<td>0.603</td>
<td>0.417</td>
<td>0.516</td>
</tr>
<tr>
<td>SVM+</td>
<td>0.378</td>
<td>0.537</td>
<td>0.641</td>
<td>0.607</td>
<td>0.405</td>
<td>0.514</td>
</tr>
<tr>
<td>LR+</td>
<td>0.369</td>
<td>0.537</td>
<td>0.725</td>
<td>0.608</td>
<td>0.408</td>
<td>0.529</td>
</tr>
<tr>
<td>Random</td>
<td>0.356</td>
<td>0.486</td>
<td>0.665</td>
<td>0.586</td>
<td>0.353</td>
<td>0.489</td>
</tr>
<tr>
<td>PR</td>
<td>0.281</td>
<td>0.428</td>
<td>0.666</td>
<td>0.520</td>
<td>0.327</td>
<td>0.445</td>
</tr>
<tr>
<td>DWFG</td>
<td>0.380</td>
<td>0.547</td>
<td>0.639</td>
<td>0.633</td>
<td>0.380</td>
<td>0.516</td>
</tr>
</tbody>
</table>

4.2 Results and Analysis

4.2.1 Comparison Results

The experiments were conducted in the 10-fold cross validation procedure, where one fold is for test and the other nine folds for training. The performance results are shown in Table 3 and 4, and the best performances in the comparisons are highlighted in bold. In the following results, we set the similarity threshold for sentence dependency $g = 0.1$ and the similarity threshold for inter-domain dependency $h = 0.8$. We will further discuss the variation of performance with different assignment of thresholds in Section 4.2.2.

From Table 3, we can see that DWFG outperforms the baseline methods in most cases in terms of both $F_1$ and ROUGE-N for document summarization. Moreover, we discover that the performances are statistically significantly improved on the MTV and ESPN domains by conducting sign test on the results, where the p values are much smaller than 0.01. In fact, we collect relatively fewer documents and corresponding tweets from MTV and ESPN compared with other domains, and thus, additional dependencies, especially cross-domain dependencies boost the performance by leveraging additional information.

In contrast to the improvements in Web document summarization, DWFG performs comparably to the simpler CRF-based methods for tweet summarization. In fact, the ground truth data are manually annotated from the perspective of readers’ interests and foci, which naturally reveals the users’ motivations for writing tweets. Therefore, the identification of important sentences from the Web document domain rarely influences the results for identifying important tweets.

4.2.2 Impact of Thresholds $\theta_g$ and $\theta_h$

In this section, we discuss the impact of thresholds $\theta_g$ and $\theta_h$ to our proposed approach. Although the proposed approach within a supervised framework can automatically learn the optimal model parameters $\theta$ based on the training instances, we still need to predefine the thresholds $\theta_g$ and $\theta_h$ to control the number of inter-domain and intra-domain dependencies in the factor graph model. Specifically, with larger $\theta_g$ or $\theta_h$, we obtain fewer dependencies, and if $\theta_g = 0$, each pair of consecutive sentences will be connected by a inter-domain factor, or if $\theta_h = 0$, all the sentences will be connected with all the tweets. To evaluate the impact of thresholds to DWFG and baseline methods (e.g., LC-CRF), we varied $\theta_g$ or $\theta_h$ from 0 to 0.1 with step length 0.01 respectively with the other threshold fixed. Due to space limitation, we only report the impact to the performance of DWFG in Figure 6(a) and (b) in terms of $F_1$, ROUGE-1, and ROUGE-2, and the performances of the baseline methods follow similar trends with different thresholds. We also plot the percentage of consecutive sentence pairs with similarity more than $\theta_g$ in Figure 6(a), and the percentage of sentence-tweet relation pairs with similarity more than $\theta_h$ in Figure 6(b).

From Figure 6(a), we can see that when $\theta_g$ increases from 0.0 to 0.5, the performance drops by 5% ~ 16% in terms of $F_1$ and ROUGE, which can be attributed to the lack of a complete view of sentence relations within the document. While with $\theta_h$ is 0.7, the performance reaches a local maximum when the retained sentence relations have a relatively higher quality. As shown in Figure 6(b), the performance of sentence identification reaches the global max-
R1(T)

R2(S)

F1, ROUGE−1, ROUGE−2
θ g

(a) Impact of θ g

0 0.2 0.4 0.6 0.8 1

20

40

60

80

100

Number of dependencies (%)

F1(S) R1(S) R2(S) F1(T) R1(T) R2(T)

liefs propagated from local factors and pair-wise dependency fac-
texts are indicated by bold font). The established inter-domain and
intra-domain and intra-domain dependencies are shown in arrows. Furthermore, be-
liefs propagated from local factors and pair-wise dependency fac-
tors in the last iteration of our inference algorithm are partly shown
with the associated variables taking values of 1 (in colored rounded
rectangles). Beliefs taking values of 0.5 indicate that the corre-
sponding factors have no preference on whether the sentences are
regarded as part of the summary or not. Beliefs taking values
greater than 0.5 convey positive attitudes, and the greater the be-
lief values, the stronger the confidence that the associated variables
should take values of 1. According to the calculated beliefs, the
summary for the social context is generated based on the selected
sentences and tweets (in bold).

We can see that the local features, e.g., statistical features, still
play a major role for social context summarization. For example,
since the most common words or phrases in the Web documents in-
clude “women”, “dead person”, “body”, “Liverpool Airport”, and
those in tweet threads include “Liverpool airport”, “Weekend At
Bernie’s”, texts that cover these words or phrases are more likely
chosen, and the probability that the relevant sentence-tweet pairs
are simultaneously selected is boosted. Moreover, various types of
relations also come into play. For example, since the last two tweets
shown in the right column form a retweet pair, the importance of
the content is evaluated more important, and thus the related sen-
tence (the fourth sentence) in the document then receives a higher
belief (0.51) of taking a positive decision. As we suggested, in the
social context summarization task, the tweet thread contributes ad-
ditional information (e.g., Weekend At Bernie’s\(^7\)) to the original
document content, which unveils the users’ interests from an alter-
native angle.

5. RELATED WORK

Web-page summarization techniques have been widely studied
for many years and various approaches have been developed. These
approaches can be either supervised or unsupervised, and also can
be generic or query-dependent. Since this paper mainly stud-
ies context summarization, we focus our literature review for ap-
paches with or without consideration of context.

Two kinds of approaches have been designed for web-page sum-
maries. Supervised and unsupervised. Traditional supervised
summarization approaches treat the summarization task as a two-
class classification problem [16, 25, 36] or a sequence labeling
problem [8, 28] at the sentence level, where each sentence is repre-
sented by a vector of features. Comparably, unsupervised methods
rely on a set of heuristic rules to develop the summarization. Web-
page summarization can also be either generic or query-dependent.
Generic summarization targets to cover the main idea of the page
while query-oriented summarization is to present the information that is
most relevant to the given queries [4, 31].

Without consideration of context, the extracted summary is com-

\(^{7}\)a 1989 American motion picture comedy, which has a similar plot
as the news story.
posed of sentences from the Web documents, and thus features from local content of a document is the key to summarization. Traditional document-oriented features can be defined either from linguistic, such as rhetorical structure [22], lexical chains [2] or statistical perspectives, such as term significance [20], sentences similarity [24] and topic detection [12]. Although document-oriented features can disclose most of the basic characteristics of summary sentences, as stated in [29], the textual information of a Web document may be scarce and diverse in topics and, moreover, contain a lot of noise.

Document-oriented features cannot fully capture the main idea of a Web document. In the past few years, some work starts to utilize various kinds of context to assist document summarization, such as external documents or cited articles [23]. User requirement is one of the most important kind of context [10, 32]. In the study of [21], user’s needs come from a set of documents selected by user, where the top content words were extracted according to their $G^2$ score and then treated as users’ interests. Hyperlinks among Web pages are another kind of context. Based on the text surrounding the hyperlink, summarization of the target Web page can be realized either by extracting the related sentences in surrounding text [1] or by extracting significant sentences from the linked Web page [9]. Similar to the hyperlink context, Sun et al. [29] utilize search-engine clickthrough data to construct the extra knowledge. In their work, Web page and query terms collected from the clickthrough data work together to decide the significance of each word in sentences for summarization. With the rapid growth of social websites, comments-oriented approach was studied, where the most important comments are selected and leveraged into sentence selection for summarization. Traditional feature-based methods and graph-based methods for summary sentence extraction have been applied for commented sentence selection [13, 33, 19], or opinionated text [26, 11, 15].

Different from previous works, we study to leverage multifaceted social media information for Web document summarization, especially social influence among users [30] and retweeting relations among messages [35]. However, we adopt a totally different approach to not only incorporate the extra knowledge extracted from microblogs, but also take full advantage of conventional techniques in single document summarization. In recent years, the rapid growth of microblogging services provide a more efficient way for information communication. Here, people can freely issue various comments on any topic they interested in. Compared with traditional tightly-coupled relationship between Web document and comments, messages from microblogs can provide more valuable information beneficial for summarization. Microblog has been widely studied in recent years. Some work focuses on investigating the characteristics of Twitter, e.g., [17], [7], [14], while some work analyzed the patterns of retweets on Twitter, influential twitter and the routines of changes of hashtags, etc., e.g., [34], [5], [3], [35]. To the best of our knowledge, little work in the literature has tried to use microblog data for Web-page summarization.

6. CONCLUSION AND FUTURE WORK

In this paper, we explore a novel problem of social context summarization and aim to utilize the mutual reinforcement between Web document and its associated social data to building a high-quality summary. In our study, the importance of each document sentence is firstly predicted by considering a series of local features of a document. At the same time, the social context relating to the Web document is associated with it, in which the significant sentences are also identified by taking advantage of various social factors. We formally define the concept of social context for Web document and propose a unified summarization approach through factor graph model. Our experiments are implemented on a set of Web documents and associated microblog messages. The experiment results prove that the proposed summarization method shows significant improvement over the baseline approaches on social context summarization task.

To systematically combine the content analysis and social behaviors represents a new and interesting direction for information retrieval. There are many future directions of this work. For example, due to the fact that not only tweets are highly associated with other tweets, users are also connected by the friendship relations, we can extend this work by establishing the connection among users and adding the dependencies between users and their

---

Figure 9: An example of a social context summary with propagated beliefs. Left column is a part of the Web document, and the right column is a portion of tweet thread. Bold texts correspond to the summary for the Web document; colored rounded rectangles indicate the beliefs propagated in the inference algorithm.
tweets. Intuitively, the influence among users will also affect the identification of important tweets, and subsequently influence the importance of sentences in Web documents.

7. *ACKNOWLEDGMENTS*

The work is supported by the Natural Science Foundation of China (No. 61073073, No. 60703059, No. 60973102), Chinese National Key Foundation Research (No. 60933013, No.61035004), National High-tech R&D Program (No. 2009AA01Z138).

8. REFERENCES