Identifying Commented Passages of Documents Using Implicit Hyperlinks

Jean-Yves Delort
University of Montpellier 2
Montpellier
France
delort@lirmm.fr

ABSTRACT

This paper addresses the issue of automatically selecting passages of blog posts using readers' comments. The problem is difficult because: (i) the textual content of blogs is often noisy, (ii) comments do not always target passages of the posts and, (iii) comments are not equally useful for identifying important passages. We have developed a system for selecting commented passages which takes as input blog posts and their comments and delivers, for each post, the sentences of the post which are the most commented and/or the most discussed. Our approach combines three steps to identify commented passages of a post. The first step is to remove the complexity of processing the contents of posts and comments using heuristics adapted to the language of the blog. The second step is to find useful comments and assigns them a degree of relevance using a model automatically built and validated by an expert. The third step is to identify important passages using relevant comments. We conducted two experiments to evaluate the usefulness and the effectiveness of our approach. The first study show that in only 50% of the posts, the most commented sentence elicited by our approach corresponds to the post extract generated using generic summarization. In the second study, human participants confirmed that, in practice, selected passages are frequently commented passages.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Natural Language Processing – Text analysis; [**Text Processing**]: Document Preparation – Hypertext/hypermedia.

General Terms

Algorithms, Experimentation

Keywords

Implicit links, passage extraction, weblogs.

HT'06, August 22–25, 2006, Odense, Denmark.

1. INTRODUCTION

Thanks to collaborative tagging systems, weblogs and other wikis, communicating on the Web is easier than ever. Weblogs are websites where authors post entries (or *posts*) that anyone can read and comment. Readers' comments often target passages of posts which may be, for instance, funny or provocative. This paper addresses the problem of identifying passages of a post which are the most commented and/or the most discussed.

Automatically identifying commented passages of a document could be useful in different applications such as document summarization. Summaries are helpful to tackle the complexity of information retrieval and processing. Methods for summarizing webpages [1, 3], websites [29] and generating snippets of search results have been proposed. These methods try to capture the gist of their targets and therefore usually focus on the author's intention. Taking into account existing comments could help produce more reader-oriented summaries. Typically, such methods could be useful to blog search engines such as technorati.[1] Another potentially interesting application of identifying commented passages is to assist moderators to detect subversive contents on weblogs.

Commenting on a post is like linking to it. Generally speaking, links often convey information about their targets. That is why, many web applications successfully rely on the context (i.e. source pages linking to a target document) to categorize [27], rank [25] or abstract web pages [1,3], as well as to support web search [26, 28]. In the case of weblogs, readers do not explicitly link to posts: either comments are inserted along with the posts or they are put together on external pages. In order to take advantage of this implicit linking, one needs to identify the exact targets of the comments.

Identifying commented passages is more specifically related to summarization by extraction and summarization by context. Summarization by extraction consists in locating passages that best characterize the content of a document [15]. Summarization by context [2] is concerned with the problem of abstracting a document from external knowledge. For example, [1, 3] have proposed methods to summarize a web page from excerpts of pages linking to it. Readers often comment on passages that are unrelated to the central ideas of the posts. The need of commenting may be triggered by an anecdote, a private joke, or

[1] http://www.technorati.com

even something unwanted by the author such as a wrong hypothesis or a misspelled name. Therefore, our problem differs from automatically summarizing a post. Combining the ideas of summarization by extraction and summarization by context is the key of our approach for selecting passages in posts using comments.

The complexity of the problem of identifying commented passages is mainly due to three reasons. First, the language used in blogs is difficult to automatically process. Posts and comments often contain numerous misspelled words, abbreviations and phonetic expressions which confuse text mining techniques. Second, comments do not always target passages of posts. This type of comments may mislead the identification of commented passages. Analyzing the types of comments is necessary to characterize *relevant comments*, i.e. comments which really target passages of posts. Third, relevant comments are not equally useful for identifying commented passages. The usefulness of a comment depends on the communicative intention and the audience targeted by the commenter. A *degree of relevance* should therefore be associated to each relevant comment.

In this article, we introduce a new approach which combines three steps to identify commented passages of a post. The first step is to remove the complexity of processing the contents of posts and comments using heuristics adapted to the language of the blog. The second step is to find useful comments and assigns them a degree of relevance using a model automatically built and validated by an expert. We have developed a method to automatically detect if a comment is relevant. Our key idea is to combine both the power of machines and human expertise to characterize relevant comments. The machine is used to present clusters of similar comments while the expert verifies whether the clusters contain only relevant or only irrelevant comments. Clusters represent prototypes of relevant and irrelevant comments and they form a model that is used to predict the relevance of comments. Finally, the third step is to select passages of the post from relevant comments. Several types of commented passages and their selection strategies are presented. We report also on the evaluation of our method and present the results of two experiments –one being quantitative and the other one involving human experts.

The paper is organized as follows. Section 2 reviews related work. Section 3 outlines our approach for identifying blog posts from readers' comments. The three following sections present with more details each of the main steps of our approach. In Section 7 we present the results of our experiments. Our conclusions and suggestions for future research are discussed in Section 8.

2. RELATED WORK

Adding comments to documents is a functionality that has long been provided by hypermedia systems [12]. Randy Trigg, Textnet's designer, distinguished fifty-one types of comment. However, Trigg's taxonomy was only intended to classify scientific research works and not any kind of comments. Comments as a special kind of annotations have also been widely studied [9, 11]. In [9], Marshall suggests several dimensions to classify annotations with respect to their audience (published - private), duration (permanent-transient) and so forth. Shipman et al. have proposed a system to identify useful passages of documents from hand-written annotations [11]. Their system aims to detect typical marks (comment, underline marks, margin bar, symbols, etc.) included in annotations of important passages.

The problem of identifying commented passages overlaps with two well-established areas of research: summarization by extraction and summarization by context. In the following, we briefly survey these issues.

2.1 Summarization by extraction

Summaries can be distinguished between abstracts and extracts [15]. An abstract contains at least some sentences that do not exist in the original document. Abstracting a document is difficult because it requires addressing complex issues such as discourse understanding and natural language generation [20]. Consequently, most approaches to automatically summarizing a document generate extracts [8, 10, 17, 18, 21, 22, 25]. An extract is made of text spans (typically, paragraphs or sentences) selected from the original document.

The most popular technique of summarization by extraction generates query-relevant summaries [8]. Query-relevant summaries are obtained by scoring sentences with respect to both statistical (such as term frequencies) and linguistic features (such as rhetorical structures, sentence location and presence of cue words [8, 10, 18]). A scoring function is used to rank the sentences with respect to a given query. The summary contains sentences with highest scores. For generic summarization, a centroid query vector is determined from the whole document. Topic-focused summaries can be generated with this approach by using different terms and weights in the query vector [8, 21, 22].

Recent works have proposed methods for identifying useful information in a document using external knowledge. Fukumoto et al. [7] have addressed the problem of selecting key passages of articles classified in topic domains. They introduce a degree of context dependency measuring the importance of a term with respect to three different contexts: the paragraphs of the article, the articles in the domain and the corpus. They define a keyword of an article by the terms with the highest context dependency in their own domain. Key passages of a document are passages containing the most important keywords. Sun et al. [13] have addressed the problem of summarizing web pages using click through data. This problem consists in finding sentences with terms frequently occurring in user queries that have lead to the page. However, rankings vary significantly on search engines even with slight changes in the query. Consequently, most pages are accessed by the same queries.

Summaries generated by extraction of sentences in the original document often suffer from lack of coherence and cohesion [24]. Moreover, multimedia documents, or documents where textual content is scarce cannot be suitably summarized by this approach.

2.2 Summarization by context

Summarization by context consists in abstracting a target document from external resources related to the target. Three important issues in web document summarization by context can be distinguished [1]: contextualization (i.e. selecting the resources that are relevant for a summary), partiality (i.e. trimming the resources that are the less informative and consistent about the target) and topicality (i.e. sorting the remaining resources and giving a good ranking to the most descriptive ones).

In [1], two methods for summarizing a webpage using the content of documents linking to it are proposed: The former uses both the content of the target and the context and the latter only uses the context and is more adapted to situations where the textual content of the target is scarce. Amitay and Paris [3] have addressed the problem of generating search engine snippets using the content of documents linking to them. They proposed a system to generate snippets from the context elements using writing conventions to determine their intrinsic quality.

The underlying hypothesis of web document summarization by context is that anchors and surrounding texts often abstract their targets [14]. Therefore, summarization by context is an effective strategy to overcome the drawbacks and limitations of summarization by extraction.

3. Discussion

Commented passages may be useful to detect what is perceived as, for example, important, original and provocative in a post. However, these passages may not reflect the central ideas of the post and cannot be identified by existing summarization techniques. To our best knowledge, identifying commented passages of a document has never been addressed in previous research. In the next section, we present our approach which combines the key ideas of summarization by extraction and summarization by context.

4. OUR APPROACH

We have developed a system which takes as input blog posts and their comments and delivers, for each post, the passages of the post which are the most commented and/or the most discussed.

Commented passages of posts could be useful to elicit the readers' favourite passages or to find controversial passages of posts. Thus, commented passages should be ranked differently according to the need of the analysis. Comments are the basis for the ranking, but it would be irrelevant to take them all into account. Indeed, a large proportion of comments is not directed to passages of the posts. Consequently many comments are irrelevant for the issue of selecting commented passages. Typical examples of irrelevant comments are spam comments or void comments such as “I love your blog, Anna”. Irrelevant comments must be discarded because they could lead to select uninteresting passages. Finding relevant comments is a complex problem because relevant comments may be quite dissimilar and because the border between relevant and irrelevant comments is sometimes fuzzy. To be suitably analysed and interpreted, blog posts and comments must be carefully pre-processed. Indeed, the specificities of the language of blogs (e.g. emoticons or phonetic expressions) may confuse text processing techniques.

Our approach breaks down the complexity of identifying commented passages into three steps: pre-processing the data, identifying relevant comments and selecting commented passages.

The first step is intended to remove the noise due to the language of blogs. For example, strings that may be misinterpreted by text processing techniques, such as misspelled terms, phonetic expressions, are located and replaced when possible by their equivalent in English. In addition, common syntactical errors are corrected using the structure of the document.

Identifying relevant comments is a complex task because relevant comments may have many different features. We have observed that the relevance of a comment highly depends on three factors: the targeted object, the target audience and the reader's communicative intention. Automatically extracting these factors is too complex a task for existing natural language processing techniques. To go through this problem, we have found basic features related to these factors that can be combined in order to characterize relevant comments. However, there is no prototype of relevant comments: Though they all target passages of the posts, they may target different audiences and have different communicative intentions.

We have designed a method to automatically discover the characteristics of relevant comments because it could be very time-consuming for a human being. Our method is based on a statistical model that is trained to distinguish between relevant and irrelevant comments. Once the model has been built, it can be used to identify relevant comments of any post. Within this model, comments are processed as feature-vectors. The feature-vectors are automatically generated from heuristics extracting the values of the basic features. The model is built thanks to an unsupervised learning algorithm because characterizations are not known *a priori* and they have to be discovered. The algorithm is intended to find clusters of comments containing either only relevant or only irrelevant comments. A human expert needs to control the outcome of the clustering to check it. If that is not the case, the expert changes the features used to represent the comments or the tuning of the algorithm and re-runs it. This process is iterative and is completed when the expert considers that the clusters are homogeneous enough. The feature-vectors of cluster centroids are considered as characterizing either typically relevant or typically irrelevant comments. The expert assigns weights to relevant clusters in order to give extra value to comments with important features during the selection of commented passages.

In the second step, the previous model is used to identify relevant comments. First, comments are turned into feature-vectors. Then, the model determines the most similar cluster of each comment by comparing the similarities of the feature-vectors with those of the centroids. Comments receive the degree of relevance of the cluster they belong to.

In the third step, passages of the comment are sorted using a ranking function adapted to the need of the analysis, which could be, for example, to select the most commented passages, or the most discussed.

To summarize, our approach has three main steps: First, the noise caused by the language of blogs is removed. Second, comments that do not target passages of posts are discarded and relevant comments are weighted. Third, commented passages are identified using a ranking function adapted to the need of the analysis. The next sections deal with the three main steps of our approach.

5. THE LANGUAGE OF BLOGS

The language of blogs is confusing for text mining techniques. Indeed, at the morphological, lexical and syntactic levels of language, blog contents contain numerous mistakes. The issue of pre-processing textual content of poor quality has been widely addressed in existing literature [1, 3]. In this section, we describe our approach to pre-processing comments and posts.

Our approach is based on the state-of-the-art data preparation techniques. First, we use a document cleaner that strips off the HTML tags of the documents. In posts and comments, numerous sentences are not finished with a dot, which may mislead the syntactic analysis carried out by the part-of-speech tagger. Therefore, we replace HTML tags suggesting the end of a sentence break by dots (e.g.
, <p>). Blogs often contain texts written in a SMS-style or misspelled terms or phonetic expressions. The exact text of these strings can be recovered with edit distances, or with dictionaries of the most current phonetic expressions. In the same vein, the frequent usage of emoticons makes difficult the issue of finding the sentence boundaries. However, dictionaries of emoticons can also be used to remove them from contents.

Then, we use a part-of-speech tagger (POS) to find the most meaningful terms in the documents [19]: adjectives, nouns and verbs. Because posts and comments often contain proper nouns and abbreviations that are tagged "unknown" by the POS, we keep all unknown terms in the document representation. We use a stop list to discard the too common and meaningless words.

6. RELEVANT COMMENTS

In order to automatically detect relevant comments, we have identified basic features of comments and developed a method to learn to characterize relevant comments. These features are related to three general dimensions of comments. In this section, we analyse these dimensions before we present our method to automatically detect relevant comments.

6.1 Comment types

Based on the analysis of a large number of blogs, we found three significant dimensions to characterize comments: the communicative intentions, the targeted audience and the targeted object. These dimensions bear resemblance with Trigg's taxonomy of commentary links [12] and with Marshall's dimensions of annotations [9]. In this subsection, we review and exemplify them.

6.1.1 Communicative intentions

A large proportion of blog comments display similar patterns of communicative intentions. It is a surprising feature given that weblogs cover a wide range of topics and that the types of readers are numerous. Typically, Trigg's taxonomy of commentary links is more detailed but also specifically adapted to comments on scientific research [12]. His taxonomy is based on the assumption that scientific research works have the same functions. He considers the following possible functions of a work: specifying context, problem posing, theory declaration, arguments and data. Then, he defines a set of possible comment types for each function. Overall, 51 comment types are listed. Most comments display the following patterns of communicative intentions:

- exhibiting feelings or opinions, e.g.:

  Hmm this post gives much food for thought! […]

- suggesting something to the author or other readers, e.g.:

  Please Jim, try to read more carefully. […]

- relating the content of post to something else, e.g.:

  Nice Blog! I've got one that similar – it's to do with Instant Messaging Security and it's here: Instant Messaging Security. Check it out Now!

- joking, e.g.:

  Ayya. You should tell that to my parents! (kidding, lol) […]

6.1.2 Targeted audience

Comments may be addressed to different people. For example, an author may reply in the comment area to a reader's question. Conversations between readers often occur in the comment areas of weblogs. Marshall proposes two criteria related to the audience of annotations [9]. At a first level she distinguishes private v. published annotations: private annotations are strictly personal whereas published annotations are meant to be read by others. At a second level, annotations can be regarded with respect to the size of the community to which they are intended; for example it can be global v. institutional v. workgroup v. personal annotations. Most comments are directed to one of the following audience:

- the post author, e.g.:

  good to see you blogging again! […]

- one or more other readers, e.g.:

  Iuridivii, your comment, while completely unintelligible, helps me to understand the complex emotions that well up inside a right-brain individual. […]

- everybody, e.g.:

  Thank you all for the comments […]

6.1.3 Targeted object

Blog comments often target information lying outside the post content. It is an important difference with annotations that are usually bound to a passage or a page. Typically, many blog comments are indirectly connected to the post subject. For example, threads of discussions between readers taking place in comment areas often deviate from the topic of the original post. Most comments are directed to one of the following objects:

- passages of the post, e.g.:

  I've been researching buying a hybrid for some time […]

- the post, e.g.:

  Stuart, please, do not stop blogging […]

- the blog, e.g.:

  If I had read your blog earlier I might have avoided some of the mistakes […]

- a previous author's post, e.g.:

  I agree with everything the previous post says […]

- the post author, e.g.:

  Hey Carolyn, I left a comment on your previous post. […]

- one or more previous comments, e.g.:

  The previous comment about our expense report […]

The previous list brings out the diversity of comment types. Relevant comments are just a subset of comments that have a peculiar targeted object but may have different targeted audiences and communicative intentions. Therefore, they are related to passages of posts at different extents. As it could be quite time-consuming for a human being to find heuristics identifying

relevant comments and determining their degrees of relevance, we have designed a method to automatically achieve this task.

6.2 Identifying relevant comments

We have developed a method to build a statistical model that automatically detects relevant comments. The key idea of our approach is to combine both the power of machines and the human expertise to characterize relevant comments. From a list of basic features related to the previous dimensions, we seek to discover characterizations of typically relevant comments. The role of the machine is to present clusters of similar comments while the role of the expert is to verify that clusters contain either only relevant or only irrelevant comments. The model is built iteratively, as long as the expert is not satisfied of the homogeneity within the clusters. However, the model is trained once-for-all, i.e. once validated the relevance of any comment can be evaluated.

Building the model requires six steps. The first step of the approach is to collect a large number of posts and their associated comments. Indeed, to be effective, the model needs to be trained with many comments. Then, basic, topic-independent features characterizing the comments are automatically extracted in order to build their feature-vectors. None of the basic features determines alone the general dimensions of the previous subsection. However, by combining them, it is possible to identify prototypes of relevant comments with good confidence. These features are connected to the content of the comment as well as its context, i.e. the post and the other comments. We have identified eight basic features:

1. The number of comments in the context containing the comment author's name.
2. The number of times authors' names of other comments in the context are quoted in the comment.
3. The number of times the post author's name is quoted in the comment.
4. The number of terms common to the post and the comment contents.
5. The number of sentences in the comment.
6. The number of links in the comment.
7. The number of unknown terms in the comment.
8. The number of terms in the list of the top-100 most frequent terms in the comments.

Third, the feature-vectors are clustered in order to exhibit categories of comments which look similar in terms of their content and their relationships to other comments and to the posts. In the fourth step, a human expert checks that each cluster contains only relevant or only irrelevant comments. This method enables the expert to just explore samples of each cluster instead of all the comment set which is interesting since clusters are far less numerous than comments. If the clusters are not homogeneous enough, the expert may decide to tune differently the clustering algorithm or to add new features to the vectors. The process is iterative and ends when the expert is satisfied of the outcome of the clustering. Once the expert has validated the model, she assigns a degree of relevance to the clusters. A higher degree should be given to clusters with comments targeting more directly passages of the posts. To summarize, our approach to building the model involves six steps:

1. Collect a large set of comments and posts
2. Find basic features to characterize relevant comments
3. Extract the features of the comments and determines their feature-vectors
4. Cluster the feature-vectors
5. Check that clusters contains comments with the same kind of relevance. If not, Go back to step 2. This step is carried out by the expert.
6. Assign a degree of relevance to each cluster. This step is carried out by the expert.

To determine the degree of relevance of a comment one first computes its feature-vector. Then, the comment is assigned the degree of relevance of the most similar cluster centroid. Once the identification of relevant comments is done, one can tackle the problem of selecting commented passages. The next section presents our method to achieve this task.

7. SELECTION

Relevant comments enable to select commented passages. However, the importance of passages depends on the need of the analysis. This section presents our method to select passages of document using comments and their associated degrees of relevance. Then, it outlines several strategies for ranking passages according to different objectives.

7.1 Selecting passages

Query-relevant summarization is a well-known technique to extract the most important passages of document. It has been widely studied and its effectiveness has been demonstrated [8, 21, 22]. It is the approach that we have chosen to select commented passages. As explained in Section 2, query-relevant summaries are obtained by scoring passages with respect to both statistical and linguistic features. To summarize a document, we first split it into passages. Then, passages are represented in the vector space model. In this model, a passage S is represented by a vector of weighted terms: $S=<s_t>_t$ where s_t corresponds to the weight of the term t. The weight of a term is given by the number of its occurrence in the text (term-frequency). Then, a scoring function is used to compare each passage with a given query $Q =<w_t>_t$. The most important passages are those with highest scores. We used the scoring function proposed in [8].

Personalizing the query enables to select different kinds of passages of the posts. For example, in generic summarization, a centroid query vector is determined from the whole document. Previous works have used this technique to generate topic-focused and user-focused summaries [21, 22]. In our approach, to select passages of a post we use the same query made up with all the terms of the post. In order to capture specific passages, we use different weighting schemes. The next subsection presents weighting schemes that enable to elicit different kinds of commented passages.

7.2 Selection strategies

This subsection outlines two weighting schemes for the terms in the query used by the scoring function to rank the passages of a post. The proposed weighting schemes are intended to select the most commented and the most discussed passages in a post. We also suggest a strategy to build weighting schemes for selecting other kinds of passages. In the following, we denote by R the set of relevant comments on a post and by d_c the degree of relevance of comment c.

7.2.1 Selecting the most discussed passages

The first step to identify the most discussed passages is to find relations between the comments. A simple heuristic for finding relations is to look for names of other readers in the content of comments. Other heuristics could be derived from existing methods to automatically extract rhetorical structures in a text [30]. A conversation graph representing the relation between comments can be drawn as shown in Figure 1.

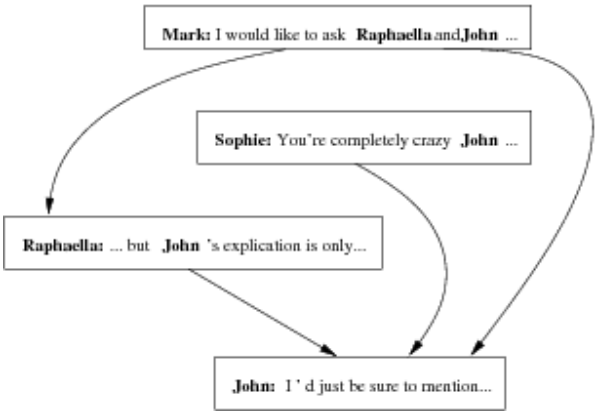

Figure 1: A conversation graph between comments.

Conversations between readers are generally initiated by the post contents. However, discussions tend to drift from the original topic. Therefore, we only trust in the K first comments to find the ideas which instigated the discussions. We define by $D \subseteq R$ the set of the K earliest comments of the discussion graphs with a size greater or equal to 2. To elicit the most discussed passages, we use the following weighting scheme for query Q:

$$w_t = \sum_{c \in D} t_c \times d_c$$

where t_c denotes the term frequency of term t in comment c.

7.2.2 Selecting the most commented passages

Selecting the most commented passages of a post is an important issue. First, it could be used to rank passages according to their popularity. Existing summarization approaches could take advantage of this ranking to reflect what the reader is likely to be interested in. Second, it could be used to abstract (to a certain extent) relevant comments on a post. Our approach for selecting the most commented passages takes into account all relevant comments. Since, relevant comments have been associated to weights, their importance in the selection process should not be equivalent. We define by the document frequency of a term t_c, the number of occurrences of t in comment c [16]. To elicit the readers' most frequent interests, we use the following weighting scheme for query Q:

$$w_t = \sum_{c \in R} t_c \times d_c$$

where t_c denotes the term frequency of term t in comment c.

7.2.3 Selecting passages using cue word.

Original passages may also be discovered from the texts following cue words or expressions such as “Nonsense, … “ or “you're completely wrong ...”. Indeed, cue words could introduce information related to passages of the posts that have something peculiar, for example, a funny joke or a provocative assertion. Important cue words could be identified thanks methods to classifying texts with respect to the sentiments they convey [4, 5, 6]. For example, Esuli et al. [6] have proposed a machine-learning algorithm that learns to recognize subjective terms that carry a positive or a negative connotation.

8. EXPERIMENTATION

This section reports the results of two experiments we conducted to evaluate our approach for selection the most commented passages of a blog post using readers’ comments. For both experiments, we have segmented the posts in sentences in order to avoid the experts examining large excerpts of text. However, methods exist to segment text in larger homogeneous blocks than sentences, for example using lexical chains [31].

The first experiment show that in only 50% of the posts, the most commented sentence elicited by our approach corresponds to the post extract generated using generic summarization. It shows that the selected passages do not convey the same information as summaries. We carried out a second experiment involving human experts in order to verify that passages selected by our approach have been frequently commented by readers.

8.1 Experimental Setup

We collected 32549 posts from blogs hosted on Blogger.com and 67441 comments associated with these posts. To gather this data, we started by building queries about politics, business and technologies that we submitted to Google. [2] We extracted the URLs of blogs hosted on Blogger.com from the result lists. Then, we fetched the latest entries posted on the blogs. Overall, we downloaded 32549 posts thanks to this process. We waited before fetching the comments because readers often comment on the posts several days later. Thus, after one week, we downloaded 67441 readers’ comments. 30.77% of posts (10017) received at least one comment and 99% of the posts have less than 20 comments. We used a part-of-speech tagger to keep only posts and comments containing at least 20% of known English terms. It enabled us to discard noisy documents or documents written in other languages.

[2] An example of a query we used: ”inurl:blogspot.com discovery NASA comments”

Table 1: Characteristics of the comments

% of comments containing a link	39.3%
% of comments containing the name of the post author	31.5%
% of comments containing the name of a comment author	3.5%
Mean number of terms in common between a post and a comment	2.5

We implemented a program to display the comments, the posts and the commented passages of the posts. It helped the expert to build the model selecting relevant comments. Table 1 summarizes preliminary results about the comments. To cluster the comments, we chose the K-means algorithm and used its implementation in the WEKA package [24]. The model is built from 1000 comments which are clustered in 40 classes. After analyzing a large sample of comments within each class and the post contents, we assigned a degree of relevance to each class. These degrees are taken into account at the second step of our approach where passages are selected using relevant comments. We assigned the weight 0 to eight classes containing only irrelevant comments and the weight 1 to twenty-nine classes containing partly relevant comments and the weight 2 to the remaining classes made up with comments directly targeting passages of their targets.

8.2 Experimental Results

Before evaluating if our approach effectively identifies the most commented passages, we want to verify that relevant comments do not always target the most representative passages in a post. If that was the case, then our approach would be similar to generic summarization. This section shows that in nearly 50% of the post, the passage identified as the most commented one is not the most representative one. Then, we describe our experiment for evaluating if the proposed approach identifies the really most commented passage and we discuss preliminary outcomes.

To elicit the most representative passage in a post, we perform generic query-relevant summarization and keep the passage with the highest score. Then, we compare it with the best passage found by our approach. Figure 2 shows the percentage of posts where the passage selected by our approach corresponds to the most representative one, given the number of relevant comments. Passages selected with our approach match the most representative passages in 50% of posts (±10%). Moreover, the percentage seems to be independent from the number of relevant comments. This suggests that there are not more commented passages when the number of relevant comments increases.

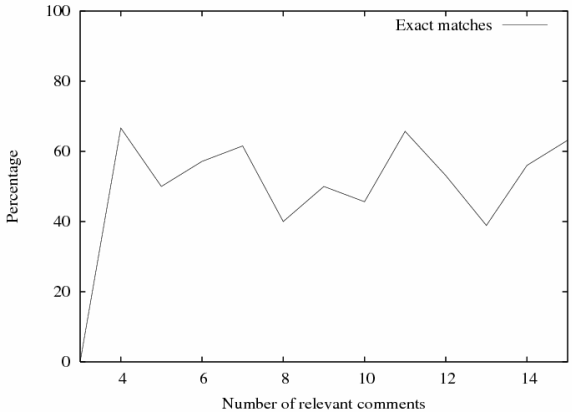

Figure 2: Percentage of posts where the most commented passage corresponds to the post summary given the number of relevant comments.

Figure 3 presents the same result but from a different perspective. The figure shows the percentage of posts where the passage selected by our approach corresponds to the most representative one given the number of sentences in the posts. Clearly, when the post contains only one sentence, then selected passage always correspond to the most representative one. The number of exact matches decreases with the size of the posts. With two-sentenced posts, the percentage is 65% whereas it goes down to 30% with posts made up with 10 sentences. Thus, the difference between our approach and generic summarization is underlined with long posts.

Figures 2 and 3 show that our approach and generic summarization rank differently their best passage in nearly 50% of the posts. This result confirms the hypothesis that comments often target passages that are out of the main topics of the posts. For example, readers' comments could focus on a passage containing a funny anecdote or on a useful information passage put in parenthesis by the author. These kinds of passages would not appear in the summary of a post but they could be discovered with our approach.

A human expert has to verify that, in practice, selected passages are the most commented passages. This task is extremely time-consuming since to evaluate a post, the expert must read it as well as all its comments. Therefore, we have designed an experimentation that enables more posts to be evaluated. The method is perfectly adapted to check that selected passages are frequently commented by readers. The principle of the approach is to make the expert choose between only a few sentences of the posts from a subset of comments. The first step consists in presenting three randomly selected relevant comments to the expert. Then, the expert is asked to choose the most commented sentence between three sentences of the post: the most descriptive sentence extracted using generic summarization, the most commented passage extracted using our approach and a randomly selected sentence.

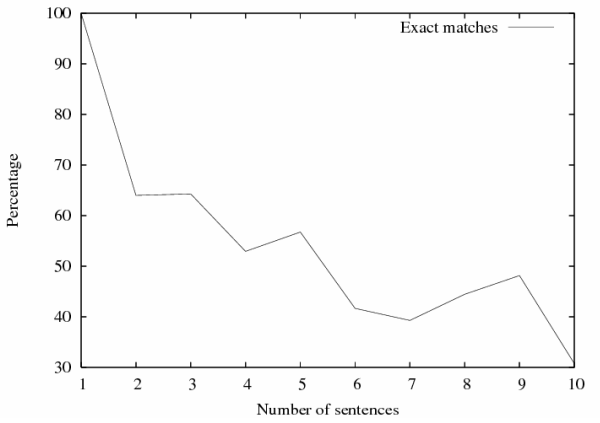

Figure 3: Percentage of posts where the most commented passage corresponds to the summary given the number of sentences.

We now illustrate how the evaluation works with an example. Figure 4 shows a post and three of its relevant comments. The comments, which have been randomly selected, are presented to the expert in order to help her to know the readers' interests in the post. We can see that the language of the comments is extremely noisy which accounts for the importance of pre-processing the data before selecting relevant comments. The two first comments address the paragraphs of the post that deal with vegetarian food while the third comment deals with US movie industry. The expert is asked to read three sentences extracted from the post and to tell which one is the most appropriate given the three previous comments. The three sentences are: the underlined sentence (the second) has been extracted as the most descriptive one; the bold sentence (the ninth) has been identified as the most commented one; the sentence with the grey background (the fifth) has been randomly selected.

We have used this methodology to evaluate the relevance of our approach for selecting the most commented passages. In order to conduct the study, we recruited adults, specifically students in Computer Science at the University of Montpellier. When the selected sentence corresponded to the most descriptive sentence, the system presented the experts with the selected sentence as well as with two other ones randomly selected in the post. Preliminary results on a sample of posts indicate that passages selected by our approach identify well commented passages.

POST: Last night Dixita, Abhishek, and a friend of theirs and I went to Byrd theater to see 'Mr. & Mrs. Smith'. I thought the movie will be an ok action movie. But to my disappointment it turned out to be stupid action movie. In the end, the movie reminded me of another movie 'Desperado' . We had earlier decided to go to a restaurant for dinner after the movie. But one person in the group (no name) broke the deal. Now I am also going to break a deal that I had made earlier with that person. Yesterday was also my first day as a vegetarian. **Now I don't know if milk, yogurt, and egg are considered vegetarian**. But I will consider them as vegetarian until unless I feel otherwise.
COMMENT 1: When have you ever listened or kept a deal with anyone else... lol Why the decision to become veggo ???? If you are vegan you cannot have milk, yoghurt, egg etc... ur still a vegetarian if you take those things.
COMMENT 2: ... So vegan is different from vegetarian. hmm ... I see . For now I will be vegetarian & let me see how it goes. And Ms. Preethi, its not 'Yoghurt' ok, it's 'Yogurt'. Yoghurt. hahahaha …
COMMENT 3: … Hey, at least we have a movie industry, ... Nicole Kidman and Russel Crowe , why did they come to Hollywood.

Figure 4: A post and a sample of three of its relevant comments

9. CONCLUSION

This paper addressed the issue of automatically selecting passages of blog posts using readers' comments. The problem is difficult because: (i) the textual content of blogs is often noisy, (ii) comments do not always target passages of the posts and, (iii) comments are not equally useful for identifying important passages.

We have developed a system for selecting commented passages which takes as input blog posts and their comments and delivers, for each post, the sentences of the post which are the most commented and/or the most discussed. Our approach combines three steps to identify commented passages of a post. The first step tackles to pre-process the contents using heuristics adapted to the type of language used on blogs. The second step identifies relevant comments and their degrees of relevance using a model automatically built and validated by an expert. The third step selects passages using relevant comments. Three selection strategies are proposed to identify: the most commented passages, the most discussed passages and passages that can be identified from the text following cue words in readers' comments, e.g. "you're completely right… ".

We conducted two experiments to evaluate the usefulness and the effectiveness of our approach. The first study show that in only 50% of the posts, the most commented sentence elicited by our approach corresponds to the post extract generated using generic summarization. In the second study, human participants confirmed that, in practice, selected passages are frequently commented passages.

In future work, we plan to propose new summarization techniques taking the popularity of passages of documents. Existing summaries are intended to reflect the authors' intentions. The approach proposed in this paper will enable to generate summaries reflecting what the reader is likely to be interested in even if the author did not stress on it.

10. ACKNOWLEDGMENTS

The author would like to express his gratitude to Marc and Jocelyne Nanard of the University of Montpellier for their advice and their help in the writing of this article. We also sincerely thank the anonymous referees for their valuable comments and assistance in improving this article.

11. REFERENCES

[1] Delort, J.-Y., Bouchon-Meunier, B. and Rifqi, M. Enhanced Web-Document Summarization Using Hyperlinks. In Proceedings of the Thirteen Conference on Hypertext and Hypermedia, pages 208-216, ACM Press, 2003.

[2] Delort, J.-Y., Bouchon-Meunier, B. and Rifqi, M. Summarization by Context, in Poster Proceedings of the Twelfth International World Wide Web Conference, 2003.

[3] Amitay, E. and Paris, C. Automatically Summarizing Web Sites – Is There A Way Around It? In Proceedings of the Ninth International Conference on Information and Knowledge Management, pages 173-179, ACM Press, 2000.

[4] Dave, D. and Lawrence, S. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of the Twelfth International World Wide Web Conference, pages 519-528, ACM Press, 2003.

[5] Pang, Bo and Lee, L. and Vaithyanathan S., Thumbs up? Sentiment Classification using Machine Learning Techniques, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 79-86, 2002.

[6] Esuli, A. and Sebastiani, F. Determining the Semantic Orientation of Terms through Gloss Classification, Proceedings of the Fourteenth International Conference on Information and Knowledge Management, pages 617-624, ACM Press, 2005.

[7] Fukumoto, F., Suzukit, Y. and Fukumoto, J. An Automatic Extraction of Key Paragraphs Based on Context Dependency. In Proceedings of the Fifth Conference on Applied Natural Language Processing, pages 291-298, 1997.

[8] Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, pages 121-128, ACM Press, 1999.

[9] Marshall, C. Toward an ecology of hypertext annotation. In Proceedings of the Ninth Conference on Hypertext and Hypermedia, pages 40-49, ACM Press, 1998.

[10] Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Celebi, A., Qi, H., Drabek, E. and Danyu Liu. Evaluation of Text Summarization in a Cross-Lingual Information Retrieval Framework. Technical Report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, June, 2002.

[11] Shipman, F.M., Price, M.N., Marshall, C.C., and Golovchinsky, G. Identifying Useful Passages in Documents based on Annotation Patterns. In Proceedings of the European Conference on Digital Libraries, pages 101-112, 2003.

[12] Trigg, R. H. A Network-Based Approach to Text Handling for the Online Scientific Community. Ph.D. Thesis, Dept. of Computer Science, University of Maryland November, 1983.

[13] Sun, J., Shen, D., Zeng, H., Yang, Q., Lu, Y., and Chen, Z. Web-page Summarization Using Clickthrough Data. In Proceedings of the 28th International Conference on Research and Development in Information Retrieval, pages 194-201, ACM Press, 2005.

[14] Menczer F. Links tell us about lexical and semantic web content. Technical report, Computer Science, abstract CS.IR/0108004, August, 2001.

[15] Sparck-Jones, K. and Galliers, J.R. Evaluating Natural Language Processing Systems: An Analysis and Review. Lecture Notes in Artificial Intelligence. No 1083. Springer, 1995.

[16] Salton, G. and McGill, M.J. Introduction to modern information retrieval, McGraw-Hill Book Company, 1983.

[17] H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, Vol. 2, No. 2, pages 159-165, April, 1958.

[18] Paice, C.D. The Automatic Generation of Literary Abstracts: An Approach Based on Identification of Self-Indicating Phrases. In O. R. Norman, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, Information Retrieval Research, London: Butterworth, 1981

[19] Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees, In Proceedings of the International Conference on New Methods in Language Processing, 1994.

[20] Spark-Jones, K. What might be in a summary? Information Retrieval 93: Von der Modellierung zur Anwendung, pages 9-26, 1993

[21] Mani I. and Bloedorn E. Machine Learning of Generic and User-Focused Summarization. In Proceedings of the Fifteenth National Conference on AI, pages 821-826, 1998.

[22] Amini, M.-R. Interactive Learning for Text Summarization. In Proceedings of PKDD'2000/MLTIA'2000 Workshop on Machine Learning and Textual Information Access, pages 44-52, 2000.

[23] Ono, K., Sumita, K., Miike, S. Abstract Generation based on Rhetorical Structure Extraction. In Proceedings of the 15th International Conference on Computational Linguistics. COLING'94, Vol. 1, pp. 344-348, 1994.

[24] Saggion, H. Automatic text summarization: past, present, and future. Tutorial of the 9th Ibero-American Conference on Artificial Intelligence. Mexico. 2004.

[25] Page, L. and Brin, S. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford Digital Library Technologies Project, 1998.

[26] Kraft, R. and Zien, J. Mining Anchor Text for Query Refinement. Proceedings of the Thirteen International World Wide Web Conference, pages 666 – 674, ACM Press, 2004.

[27] Attardi, G. and Gull, A. Automatic Web Page Categorization by Link and Context Analysis. In Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, pages 105-119, 1999.

[28] El-Beltagy, S.R. and Hall, W. and Roure , D. and Carr, L. Linking in Context. In Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia, pages 151-160, ACM Press, 2001.

[29] Zhang, Y. and Zincir-Heywood, N. and Milios, E. World Wide Web Site Summarization, Technical Report, Faculty of Computer Science, Salhousie University, April, 2002.

[30] Marcu, D. A decision-based approach to rhetorical parsing. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 365-372, 1999.

[31] Stokes, N. Carthy, J. Smeaton, A.F. Segmenting Broadcast News Streams using Lexical Chains, In Proceedings of STarting AI Researchers Symposium, pages 145-154, 2002.