2008－GN－68（6）
2008／5／23

Identifying Unusual Blog Entries: A Behavioral Approach Analyzing Blogger's "Friends"

Roland Hou-Yin Hui Akihiro Miyata Harumi Kawashima Hidenori Okuda
NTT Cyber Solutions Laboratories, NTT Corporation
1-1, Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847, Japan
<roland.hui, miyata.akihiro, kawashima.harumi, okuda.hidenori>@lab.ntt.co.jp

Abstract

Unusual events in life are often the events we treasure, find interesting and consequently share with friends. We present a behavioral method to identify unusual blog entries within a blog by evaluating comment senders' behaviors. Reader Feedback Vector (RFV) represents each blog entry with a comment sender properties vector and examines each vector for any indication of abnormal comment sender behaviors. An Intimacy Score, which measures the "friendship" level between each comment sender and blogger, is used to weigh each comment sender's properties. This ensures that comment senders who are deemed more "intimate" with the blogger are evaluated more importantly. RFV is subjected to a human-based analysis which asks each test subject to compare the results from RFV to the results from a text-based approach. A second similar test evaluates the significance of Intimacy Score by comparing RFV with and without Intimacy Score (RFV-IS vs. RFV). RFV-IS performed better, by returning more unusual blog entries than its counterpart.

Keywords: Unusual events, Blogs, Behavior, Intimacy, Friends, Bloggers

1. Introduction

Life can be thought of as a "mash up" of experiences from countless events in the past, present and future. Most of these events are quite ordinary and even a little bit mundane from the eyes of others. However, there are times when one experiences an unusual event; events where one undoubtedly remembers and shares with friends. The definition of unusual events cannot be clearly defined as it differs from person to person; some may think that their high school graduation is unusual, whereas others do no. Whatever the actual event may be, these unusual events are the ones where one reminisces about years after it takes place.

LifeLog is an attempt to create one's profile by recording and storing every aspect of one's life (including blogs) in an electronic database. Thus, the unusual events of one's life are extremely significant. Although not restricted to LifeLogs, we propose a method called the Reader Feedback Vector with Intimacy Score (RFV-IS) to select and present blog entries with highly unusual events within one blog. In this way, LifeLog's can use these blog entries to summarize one's life's unusual events. Similarly, blog readers also benefit from this technology; being presented with an alternative to quickly and effectively better understand the blogger.

The rest of this paper is organized as follows. In Section 2, some background and related works are discussed. Section 3 presents our goal as well as the approach used to achieve this goal. An evaluation of this approach is discussed in Section 4. Finally, the conclusion and future works are offered in Section 5.

2. Related Works and Background

There are two main types of work related to the identification of unusual blog entries. Classification or ranking of blog entries is one related type of work. A second type of work involves the analysis of blog readers by evaluating blog feedback mechanisms such as comments and trackbacks.

2.1. Classification or Ranking of Blog Entries

This type of research focuses on differentiating blog entries from one another. Ni et al. gauges several clustering techniques such as Naïve Bayes Classifier (NB) and Support Vector Machine (SVM) to automatically classify blog entries as either 'informative' or 'affective' articles. Ni et al. defines 'informative' articles as "news that is similar to the news on traditional news websites, technical descriptions and commonsense knowledge." The 'affective' articles are defined as "Diaries about

personal affairs and self-feelings or self-emotions" [1].

The following two researches are less related to our work due to the classification or ranking of blogs relative to other blogs in the blogosphere. Therefore, the domain in which the following two technologies operate on is much wider. Nakajima et al. identifies important bloggers, namely Agitators, "[One who] stimulates the discussion in a blog thread so that it becomes more active" and Summarizers, "[One who] summarizes a hot blog thread by referring to many other entries," by taking several approaches using a combination of links, time and text [2]. Fujimura et al. offers a ranking algorithm that assigns a score to a given blog entry by taking into account the blogger's previous blog entries. In this way, a higher score can be "assigned to the blog entries submitted by a good blogger but not yet linked to by any other blogs" [3]. Classification of blog entries is important in the identification of unusual blog entries. However, these classification methods are not suitable for the identification of unusual blog entries on a "per blog" level.

2.2. Analysis of Blog Feedback Mechanisms

Blog feedback mechanisms – ways in which readers can interact with the blog – such as comments or trackbacks are usually excluded from existing blog research as noted by Hu et al. Their research proposes to use both the text from a given blog entry and its blog comments to summarize the entry [4]. Beibei et al. enhances blog entry clustering by using both the author and reader comments [5]. Miyata et al. considers the number of blog comments, trackbacks and other blog feedback mechanisms in their blog search algorithm [6]. Although our approach also makes use of blog feedback mechanisms, these works are not suitable for the identification of unusual blog entries on a "per blog" level.

3. Proposal

3.1 Main Goal

Blogs consists of a number of blog entries; some of which can be considered unusual while most would be considered quite ordinary. The identification of blog entries containing unusual events, also known as unusual blog entries, can be used in various fields. One such example is in Lifelogs, where these unusual blog entries can aid in the summarization of a blogger's unusual event history. Therefore, our goal is to automatically identify blog entries which contain unusual or unique events. The definition of this type of blog entry is one that describes an event drastically different from other events described in blog entries within the same blog. In effect, each blog has its own specific definition of an unusual blog entry due to the uniqueness of each blogger. All unusual events can be classified in one of the following groups (1) Distinct unusual events (2) Indistinct unusual events.

Distinct unusual events are defined as events that can be classified as unusual without the need to be acquainted with the blogger. Indistinct unusual events are vaguer unusual events which require a deeper understanding of the blogger before being able to judge such blog entries. For example, within the personal diary blog domain, the two unusual event groupings would contain events such as the ones listed in Table 1.

Table 1: Examples of unusual events in the personal diary blog domain

Distinct Unusual Events:
Life milestones – graduation, marriage, birth etc. **Tragic events** – accidents, deaths etc.
Indistinct Unusual Events:
Winning a sports game – Blogger's team has not won a game in two seasons. **Going to a festival** – Blogger has not been to a festival in ten years.

From the examples in Table 1, it is important to note that these events may differ in terms of unusualness from blogger to blogger.

3.2 Problems with Existing Approaches

Due to the obvious abundance of text in most blog entries, current blog analysis technologies focus mainly on text-based methods. Therefore, a preliminary survey using a simple keyword search algorithm was performed on several blogs to identify unusual blog entries. However, there was one major flaw in this text-based approach due to the unique definition of an unusual blog entry. Keyword searching proved impractical; while distinct unusual events could be identified using a brute force

keyword search, indistinct unusual events proved to be unidentifiable due to its blogger specific properties. A keyword representing an unusual event for Blogger A may represent a very usual event for Blogger B.

A second approach using text-based clustering was tested. However, blog entries relying on the heavy use of pictures was one major problem and decreased the overall effectiveness of this text-based clustering approach significantly. An evaluation of this text-based clustering approach is offered in Section 4.

3.3 Proposed Method

To identify unusual blog entries, our proposed method, Reader Feedback Vector with Intimacy Score (RFV-IS), evaluates blog comment senders' behaviors. By evaluating each comment senders' behavior, RFV-IS essentially extrapolates the thoughts and feelings that each comment sender has towards a blog entry. Furthermore, this approach is based on the wisdom of crowd idea, where all blog comment senders' reactions, exemplified in their comments, are considered and evaluated accordingly.

This method takes advantage of two assumptions based on highly correlated observed behaviors. (1) An unusual blog entry within a blog provokes unusual or unique commenting behaviors from its blog comment senders. (2) Blog comment senders who post comments more frequently are considered more in-tune with the blogger and thus have a better understanding of the blogger. From these assumptions, Reader Feedback Vector with Intimacy Score represents each blog entry as a multi-dimensional vector, where each cell consists of one comment sender's comment properties within the blog entry in question. Specifically, each cell contains the following comment sender's properties:

- Number of comments posted.
- Average length of comment.
- Average number of pictograms.
- Average number of out-links.

Aside from these properties, another value named the Intimacy Score is used to represent the level of intimacy between a comment sender and the blogger. This value is evaluated for every blog comment sender of a blog and is based on the number of unique blog entries each sender has commented on. The Intimacy Score, is used to weigh the importance of each comment sender; allowing each comment sender's properties to be evaluated accordingly.

With each blog entry represented as a vector of blog comment sender properties, RFV-IS applies data analysis techniques to detect blog entries with significant property differences. Clustering is one such data analysis technique and in short, is "the process of organizing objects into groups whose members are similar in some way" [7]. While a more in-depth definition of clustering is out of the scope of this paper, clustering serves as a practical tool for the detection of unusual blog entries because it handles multi-dimensional data effectively. It uses the multi-dimensional vectors and Intimacy Scores for the inputs and clustering weights (the importance placed on a given attribute when clustering) respectively.

From the definition of clustering, small clusters are detected and labeled as unusual because of their comment sender property differences relative to neighboring clusters.

4. Evaluation

To measure the effectiveness of Reader Feedback Vector with Intimacy Score, we perform two human-based evaluations. First, we compare a plain version of Reader Feedback Vector without Intimacy Score (RFV) with a text-based clustering method by comparing the unusual blog entries returned from each method. Second, we analyze the importance of Intimacy Score by comparing two versions of RFV, one with and one without Intimacy Score. Again, we compare the blog entries generated from RFV-IS to the blog entries generated by RFV.

4.1 Evaluation Details

In order to evaluate the effectiveness of each method, listed below, this human-based evaluation presents unusual blog entries to test subjects and asks each subject to answer our evaluation form.

1. Reader Feedback Vector with Intimacy Score.
2. Reader Feedback Vector without Intimacy Score.
3. Text-based Clustering.

Initially, two blogs, Blog A and Blog B, are chosen from a set of personal diary type blogs where unusual events are defined in Table 1 under the

headings Distinct and Indistinct Unusual Events. Next, each of our ten test subjects is assigned to evaluate either Blog A or Blog B and is given 150 of the latest blog entries from the assigned blog. These blog entries are used by each test subject to acquire a better understanding of the blogger, such as the writing style of the blogger, the tone of the blogger or the type of topics written by the blogger. Once this step is complete, each subject is presented with three unusual blog entries from each of the three methods, for a total of nine blog entries, and is presented with the following statement:

"This blog entry contains an unusual event relative to other events contained within this blog." (1)

Where the question is answered on a six-grade scale '1' to '6' with '1' representing total disagreement and '6' representing total agreement.

The next section describes how each of the three methods (1-3) identifies unusual blog entries. This is followed by a section describing the procedures taken to evaluate and validate the test results from each test subject.

4.1.1 Reader Feedback Vector with Intimacy Score

For the evaluation version of RFV-IS, each blog entry is represented as a multi-dimensional comment sender properties vector as described in Section 3. The clustering technique chosen for this evaluation is K-means clustering algorithm. Each comment sender's Intimacy Score is calculated as the number of unique entries commented. These comment senders are then ordered in terms of Intimacy Score and ranked from highest to lowest with the bottom 20% of comment senders being assigned a clustering weight of 0.

To obtain the three unusual blog entries used in the evaluation, three single blog entry clusters are chosen where the distance between each cluster and its neighboring clusters are maximized.

4.1.2 Reader Feedback Vector without Intimacy Score

The evaluation version of RFV operates in exactly the same way as RFV-IS except obviously, it does not calculate an Intimacy Score for each comment sender. As a result, the K-means clustering algorithm effectively clusters the blog entries based on the properties of every comment sender. Again, the distance between these clusters and neighboring clusters are maximized.

4.1.3 Text-based Clustering

In order to determine unusual blog entries using a text-based clustering approach, the following assumption is used, "Unusual blog entries possess many words that are uncommon in other blog entries due to a difference in topic or event." Based on this assumption, the Text-based Clustering method represents each blog entry as a multi-dimensional text vector where each cell value represents the frequency of a word contained in the blog. Specifically, we calculate the Term Frequency – Inverse Document Frequency (TF-IDF) for each word. Using these blog entries, represented by TF-IDF word vectors, as input sources of K-means clustering algorithm, three unusual blog entries are once again obtained.

4.1.4 Test Result Validation Procedure

As mentioned previously, each subject is assigned to either Blog A or Blog B. Each subject is then given nine unusual blog entries from the assigned blog, three blog entries generated from each of the three methods. These unusual blog entries are identical for each test subject assigned with the same blog. They are, however, viewed in random order to reduce order bias from each test subject.

To evaluate the test results from each test subject. The average of the six-grade scale value used to represent the agreement level with statement (1) is taken per method per test subject. As a result, each test subject has a single value, referred to as the agreement average, to represent his or her general agreement with statement (1) for each method. The agreement average of each method is then compared amongst the three methods to judge the overall effectiveness of each method; the higher the agreement average, the more effective the method. Similarly, taking the average of each method amongst all the test subjects eliminates test subject specific bias and returns a more accurate analysis of the overall effectiveness of each method.

4.2 Evaluation Results

The test results are summarized in Table 2 and Table 3. Table 4 combines the total agreement

average of Blog A and Blog B. (Refer to the Appendix for a breakdown of each subject's agreement average.)

Table 2: Evaluation results from Blog A

Method Name	Agreement Avg. (AA)
RFV-IS	4.60
RFV	3.33
Text-based Clustering	3.13

Table 3: Evaluation results from Blog B

Method Name	Agreement Avg. (AA)
RFV-IS	4.53
RFV	3.40
Text-based Clustering	3.00

Table 4: Combined evaluation results from Blog A and B

Method Name	Agreement Avg. (AA)
RFV-IS	4.57
RFV	3.37
Text-based Clustering	3.07

4.2.1 Reader Feedback Vector without Intimacy Score vs. Text-Based Clustering

This test compares Reader Feedback Vector without Intimacy Score against the Text-based Clustering method. Looking at Table 4, an initial comparison between the agreement averages from the RFV (AA: 3.37) and the Text-based Clustering methods (AA: 3.07) indicates that RFV is slightly more effective than Text-based Clustering in terms of identifying unusual blog entries. However, a Wilcoxon's paired signed ranked test between the two methods returns a p-value of 0.07488, which is greater than the 5% level. Therefore, RFV and Text-based Clustering Method cannot be proven as "statistically significant."

4.2.2 Reader Feedback Vector with and without Intimacy Score

To determine the significance of Intimacy Score in the identification of unusual blog entries, a test between RFV-IS and RFV is performed. Looking at Table 4, the agreement averages of RFV-IS (AA: 4.57) is much greater than the averages returned from RFV (AA: 3.37). Moreover, a Wilcoxon's test returns a p-value of 0.0104 which is much lower than the 5% level. Therefore, it can be said that these two methods are "statistically significant," proving that Intimacy Score is a significant enhancement to RFV.

For completion, a second test between RFV-IS and the Text-based Clustering approach is performed. Again, RFV-IS performs better, with a p-value of 0.00764.

4.3 Evaluation Discussion

From the evaluation results in Section 4.2, we can see that the differences between the behavioral and text-based method results are not very significant. In fact, it is "statistically insignificant" due to p-value > 0.05. However, with the addition of Intimacy Score, we witness "statistically significant" differences between RFV-IS and its counterparts. This is due to Intimacy Score's ability to treat every comment sender differently, placing higher importance on behaviors from comment senders who have a better understanding of the blogger. An examination of the returned unusual blog entries by RFV-IS consists of both distinct and indistinct unusual blog entries.

5. Conclusion

We provided a behavioral approach using Reader Feedback Vector with Intimacy Score to identify unusual blog entries effectively. This approach essentially extrapolates the thoughts and feelings comment senders hold towards a given blog entry. Through the detection of unusual comment sender behavior, this approach was able to identify both distinct and indistinct unusual blog entries. In addition, the Intimacy Score gives evaluation preference to "friends" of the blogger due to the assumption that "friends" are more knowledgeable regarding the blogger than "non-friends."

There are various situations where the identification of unusual blog entries could prove practical. One such example is in lifelogs, where RFV-IS can be used to identify unusual blog entries for the summarization of one's life events. The application of RFV-IS in blog search engine technologies would be beneficial; finding unusual blog entries to be either returned or filtered. RFV-IS can alternatively be applied in the Social Network Services (SNS) domain to build unusual or interesting profiles of each user based on friends' behaviors extrapolated from SNS specific feedback mechanisms.

In the future, we would like to enhance RFV-IS with the addition of comment sender text attributes such as the usage frequency of particular "important" words within a given blog.

6. References

[1] Ni, X., Xue, G., Ling, X., Yu, Y., Yang, Q. Exploring in the Weblog Space by Detecting Informative and Affective Articles. In *Proceedings of the 16th International Conference on World Wide Web,* 2007

[2] Nakajima, S., Tatemura, J., Hino, Y., Hara, Y., Tanaka, K. Discovering Important Bloggers based on Analyzing Blog Threads. In *Frontiers of WWW Research and Development,* 2006

[3] Fujimura, K., Inoue, T., Sugisaki, M. The EigenRumor Algorithm for Ranking Blogs. In *WWW 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics,* 2005

[4] Hu, M., Sun, A., Lim, E. Comments-Oriented Blog Summarization by Sentence Extraction. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management,* 2007

[5] Beibei, L., Shuting, X., and Jun. Z. Enhance clustering blog documents by utilizing author/reader comments. In *Proceedings of the 45th Annual Southeast Regional Conference*, 2007

[6] Miyata, A., Matsuoka H., Okano, S., Yamada, S., Ishiuchi, S., Arakawa, N. and Kato, Y.: Blog Search Method Based on Analysis of Response to a Blog Entry. In *IPSJ Journal Vol.48, No.12,* 2007, 4041-4050

[7] Matteo, M. A Tutorial on Clustering Algorithms,: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html

Appendix

Blog A	Agreement Average		
Subject ID	**RFV-IS**	**RFV**	**Text-based Clustering**
Subject 1	4.67	3.00	2.67
Subject 2	4.67	4.33	3.67
Subject 3	5.00	3.00	3.33
Subject 4	3.33	3.67	3.67
Subject 5	5.33	2.67	2.33

Blog B	Agreement Average		
Subject ID	**RFV-IS**	**RFV**	**Text-based Clustering**
Subject 6	4.67	3.00	3.33
Subject 7	4.33	2.67	2.00
Subject 8	5.00	4.33	3.67
Subject 9	4.67	3.33	3.33
Subject 10	4.00	3.67	2.67