Big Community Data before World Wide Web Era

Tomoya Iwakura[†], **Tetsuro Takahashi**[†], **Akihiro Ohtani**[‡], **Kunio Matsui**[‡]
[†]Fujitsu Laboratories Ltd. [†]NIFTY Corporation
[†]{iwakura.tomoya, takahashi.tet}@jp.fujitsu.com
[‡]{ohtani.akihiro, matsui.kunio}@nifty.co.jp

Abstract

This paper introduces the NIFTY-Serve corpus, a large data archive collected from Japanese discussion forums that operated via a Bulletin Board System (BBS) between 1987 and 2006. This corpus can be used in Artificial Intelligence researches such as Natural Language Processing, Community Analysis, and so on. The NIFTY-Serve corpus differs from data on WWW in three ways; (1) essentially spam- and duplication-free because of strict data collection procedures, (2) historic user-generated data before WWW, and (3) a complete data set because the service now shut down. We also introduce some examples of use of the corpus. We plan to release this corpus to research institutes for research purpose. In order to use this corpus, please email to forum-corpus@list.nifty.co.jp.

1 Introduction

The online data on World Wide Web (WWW), such as Twitter [1], Facebook [2], and so on, are widely used for the research of Artificial Intelligence (AI). In this paper, we introduce a new corpus, NIFTY-Serve corpus, which includes a big community data before WWW.

NIFTY-Serve service was carried on from 1987 to 2006 in Japan and a big social network that had about 500 thousand users and 40 million postings. The data consists of not only texts but also movies, music files, programs, and so on. From the NIFTY-Serve data, we extracted texts by a CSV format as NIFTY-Serve corpus. In total, 27,943 text files mainly written in Japanese have been extracted. The total size of the text files is about 35 GB as of May 2014. Each text file corresponds to a bulletin board and a post of a bulletin board is a text annotated with metadata like the user who posted, the posting date, related posts, and so on. There are archives of BBS for English[3][4], however, to the best of our knowledge, this is the first release of a large amount of Japanese BBS data before WWW.

In this paper, we first introduce characteristics of NIFTY-Serve data in Section 2. Then, we describe NIFTY-Serve corpus in Section 3 and some of the examples of use of the NIFTY-Serve corpus in Section 4. Finally, we briefly introduce its disclosure condition in Section 5.

2 Characteristics of NIFTY-Serve Data

The NIFTY-Serve data has the following three main characteristics different from data on WWW. These characteristics make the NIFTY-Serve data worth to use for researchers.

2.1 Well Identified Users and Quality Contents

Users were required to register their credit cards or bank accounts for the use of NIFTY-Serve. In addition, in order to identify users, NIFTY-Serve sent letters to users for confirmation on a routine basis.

[1] https://twitter.com/
[2] https://www.facebook.com/
[3] http://www.rcat.com/fido_public/
[4] http://bbslist.textfiles.com/support/sources.html

Proceedings of the 12th Workshop on Asian Language Resources,
pages 68–72, Osaka, Japan, December 12 2016.

Figure 1: An example of a bulletin board of NIFTY-Serve.

P-ID	R-ID	NIFTY-ID	Handle ID	Post-Title	Posting-Date	Main-Content
54832	54830	ID01148299	ID01148284.handle_alpha-number_.-708377626.199301051731 00	Re: By the way	960506711	Hello Mr. Smith. I have just copied the data for John. Tom

If there were letters that were not delivered to users, the accounts of the users were deleted. As a result, the users were well identified and there were essentially no user duplication and spam users.

In addition, contents of the NIFTY-Serve kept quality because the usage fee of a user was charged by how long the user connected to the NIFTY-Serve. When they posted their comments to the NIFTY-Serve, in order to avoid waste of money, users well considered what they were posting. In addition, administrators managed posts from users. As a result, NIFTY-Serve maintained better quality and there are fewer meaningless posts like just greeting that can be seen often on WWW.

2.2 Historic User Generated Data

The NIFTY-Serve service was carried on from 1987 to 2006. Therefore, the NIFTY-Serve data includes a large amount of text data written in Japanese before WWW became popular. One of the prominent examples is texts related to the Great Hanshin/Awaji Earthquake data that is an important record that includes how people acted in the disaster on online communities.

2.3 Complete Online Communities Data

NIFTY-Serve already finished its service. The complete data can be seen as a record of an online community service from beginning to end. NIFTY-Serve data includes both friendship-based communities such as Facebook and content-oriented communities such as YouTube[5] as mentioned in (Asatani et al., 2013). Therefore, we can see NIFTY-Serve as a data set that includes the whole lives of different types of online communities simultaneously.

3 Data Format

The NIFTY-Serve data includes personal information. Therefore, in addition to the original data, to reduce the risk of the leakage of the personal information as much as possible, we have prepared an anonymized corpus for the NIFTY-Serve data.

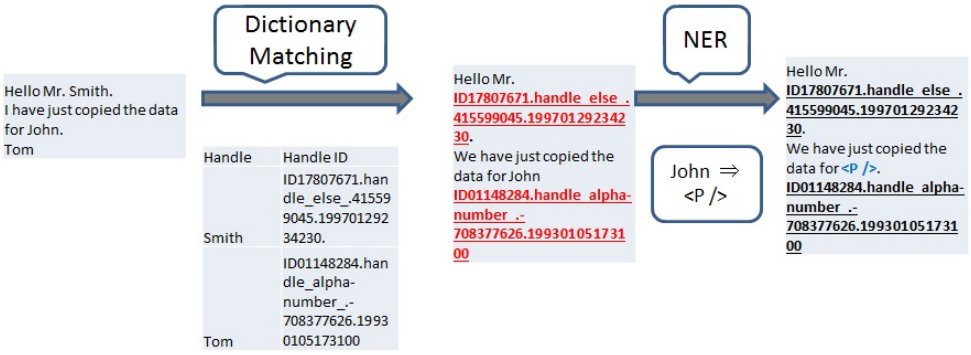
3.1 Original Data

We first introduce the original data format. Each text file corresponds to a BBS and the file name of the text file is the title of the BBS. In each text file, a post is represented by a CSV format. Figure 1 shows an example of a self-produced post. The first row indicates items of each post. The following are descriptions of the items.

- Post-ID (P-ID): A unique integer for a post.
- Response-ID (R-ID): The Post-ID of the post that is replied by a post. If Response-ID is 0, the post is not a reply to any other posts.
- NIFTY-ID: This corresponds to a user account that consists of 8 digits integer numbers beginning with “ID”.
- Handle ID: The id of a user in a BBS.
- Post-Title: The title of a post.

[5] http://www.youtube.com/

Figure 2: A process for generating an anonymized corpus.



- Posting-Date: This is the date that indicates when a post was published. The format is yymmddhhmmss, where "yy" indicates the last two digits of the dominical year, "mm" indicates a month, "dd" indicates a day, "hh" indicates hour, "mm" indicates a minute, and "ss" indicates a second.
- Main-Content: The text of a post.

3.2 Anonymized Corpus

This section introduces an anonymized version of NIFTY-Serve corpus. This corpus is mainly used for situations that users analyze as original text information as possible except sensitive personal information. In this corpus, personal names in the Main-content of each post were removed. Figure 2 shows an example of a generation of an anonymized data. First, we use a dictionary that includes pairs of Handles and "Handle ID". If there are words corresponding to handles the words are replaced with their corresponding "Handle ID".

To remove person names not included in the dictionary, we used a Japanese Named Entity (NE) recognizer (Iwakura, 2011). After identifying personal information in the text of a post, we replace the personal information by some meaningless symbols and the other words are still remained. The "... data for ⟨P/⟩. ..." in Figure 2 is an anonymized part by NER.

4 Examples of Use of NIFTY-Serve Data

This section describes four examples of the use of NIFTY-Serve data. We believe not only these examples, but also the other uses would be found by releasing this corpus.

4.1 Analysis of Changes of Word Usage

One of the examples is an analysis how use of words changes. We experimentally extracted emoticons by regular expressions. From the extraction results, we saw the following. From 1990 to 1998, a face mark "(^^)" that indicates happiness or joy was frequently used. However, around 2000, a face mark "m(_ _)m" that indicates apology or request was frequently used. This is a simple analysis, however, there may be possible to indicate something because the NIFTY-Serve data includes few meaningless information in our observation. In order to analyze real meanings of the use of face marks, one of the options is to collaborate with the other domain experts like sociologists or linguists.

4.2 Comparative Investigation of Online Communities

NIFTY-Serve data includes texts related to some historic events like disasters. One of the prominent examples is texts related to the Great Hanshin/Awaji Earthquake data. In Japan, we had the greatest earthquake, the Great East Japan Earthquake, on March 11th, 2011 since the Great Hanshin/Awaji Earthquake on January, 1995. At the Great East Japan Earthquake, services on WWW like Twitter are used for announcing information about emerging evacuation area, food support, the confirmation of the

safety of disaster victims, and so on. These data are being used for researches like how users behave in such disaster (Inui et al., 2013).

To the best of our knowledge, texts in NIFTY-Serve are one of the biggest text data set related to disasters other than the earthquake on March 11th. By analyzing the both data with NLP technologies, we expect to find knowledge for the prevention of disaster.

4.3 Benchmark Data

NIFTY-Serve data includes some quality metadata and posts with such metadata can be used as benchmark data. One of the examples is an author identification task (Inoue and Yamana, 2012) by using posts annotated with user information. We think this data set is one of the most quality data for author identification tasks due to the characteristic described in Section 2.1. We expect to find the other uses by releasing this data to research communities.

4.4 Discovering Missing Data

NIFTY-Serve data also includes data other than texts such as multimedia data, programs, and so on. Therefore, we expect to discover lost multimedia data from the NIFTY-Serve data. For example, when we were converting the original NIFTY-Serve data to the NIFTY-Serve data set, we found an archive of computer viruses that is difficult to find today. We can also use the data as a source for discovering predominant multimedia data and programs before WWW became popular. Such old data discovery may help to know what happened before WWW became popular and the analysis of such data may contribute new discoveries.[6]

4.5 Online Community Analysis

NIFTY-Serve data includes a large amount of community data based on bulletin board posts. The data includes information such as when a community began and ended, how many posts and user each community had, and so on. The information would be helpful for the analysis of online communities as described in (Asatani et al., 2013). In addition, by comparing the community of NIFTY-Serve corpus with the current communities on WWW, we could find a common phenomenon shared with online communities.

5 Disclosure Condition

We cannot openly release the NIFTY-Serve corpus because of the inclusion of personal information and the condition of the contract with users. Therefore, we release this corpus for research purpose in research institutes under one of some contract types for the release of the corpus. If users make a contract with us, the users can use the NIFTY-Serve corpus by one of the formats described in Section 3. Please email to `forum-corpus@list.nifty.co.jp` for more detail of this corpus.

6 Conclusion

We have introduced the NIFTY-Serve corpus and some examples of use of the corpus. NIFTY-Serve corpus has some prominent characteristics that are different from data on WWW such as well identified users, quality contents, and so on. We also introduced examples of the use of the corpus like an analysis of the use of words. In the future, in order to contribute research communities, we plan to release this data to users by making a contract with us.

References

Kimitaka Asatani, Fujio Toriumi, Hirotada Ohashi, Mitsuteru Tashiro, and Ryuichi Suzuki. 2013. Prosperity and decline of online communities. In *PRIMA*, pages 396–404.

[6]Unfortunately, a worker unintentionally decompressed the archive and some machines were infected by the viruses. However, a virus software detected the viruses and killed the viruses.

Masatobu Inoue and Hayato Yamana. 2012. Authorship attribution method using n-gram part-of-speech tagger: Evaluation of robustness in topic-independence (in japanese). *DBSJ Journal*, 10(3):7–12.

Kentaro Inui, Hideto Kazawa, Graham Neubig, and Masao Utiyama. 2013. *Workshop on Language Processing and Crisis Information 2013*. IJCNLP 2013.

Tomoya Iwakura. 2011. A named entity recognition method using rules acquired from unlabeled data. In *Proc. of RANLP'11*, pages 170–177.