# Visualization of Statistics from MEDLINE®

Jongwoo Kim and George R. Thoma
Lister Hill National Center for Biomedical Communications.
National Library of Medicine
Bethesda, USA
jongkim@mail.nih.gov

Paul LoBuglio
Department of Computer Science
University of Toronto
Toronto, Canada

***Abstract*—We propose a system to visualize statistics collected from NLM's MEDLINE® database that contains citations related to biomedical journal articles. The system extracts information from author affiliations in the articles such as organization, city, state, country, etc., categorizes the articles into several groups using the information, collects statistics such as the number of articles published per country each year, etc., and displays the statistics through a Web site using tables and choropleth maps. Hidden Markov Model (HMM) and statistics are used to extract the information from the affiliations, and Google Map API, JSON, JavaScript and other APIs are used for the development of the site.**

***Keywords-MEDLINE; statistics; affiliation; visualization***

## I. INTRODUCTION

The U.S. National Library of Medicine (NLM) maintains MEDLINE, a bibliographic database containing over 22 million citations related to the biomedical journal literature [1]. The number of citations in MEDLINE is rapidly increasing every year and NLM collects some statistics each year. However, there are no detailed statistics such as the number of citations published per country each year, the number of citations per organization each year, or the number of citations that received grants from NIH per country each year, etc. In addition, there is no citation field for country, organization, etc. in the existing citations. Since the number of publications can be a measure of active research in countries or organizations, these statistics may be useful information for researchers, students, or granting organizations. We therefore propose a system to collect and visualize the statistics through a Web site.

## II. SYSTEM OVERVIEW

The overview of the system is shown in Fig. 1. "Information Extraction Program" retrieves XML files of articles from MEDLINE, extracts author information from affiliations in the files using machine learning algorithms, computes statistics based on this information, and saves the statistics in the system database. Users can view the statistics through the Web site we propose.

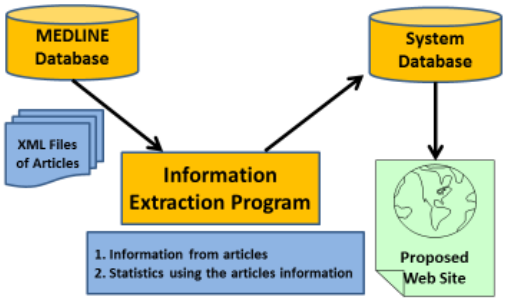


Figure 1. Workflow of the proposed system.

## III. AUTHORS' AFFILIATIONS

There are several types of affiliations in MEDLINE. Some affiliations show just organization names and others show full mailing addresses of their organizations. We extract seven different labels from affiliations [Organization (Org), City, State, Postal Code (PC), Country, Email, and Other] as shown in Table I.

## IV. PROPOSED APPROACH FOR INFORMATION EXTRACTION

1. **Word standardization**. All non-standard words are replaced with our own standard words. For example, "P R China", "PR of China", etc. are converted to "China".
2. **Postal code detection**. Every country has its own format for postal codes. We therefore search the codes of several countries using Google search [2], collect the codes for 121 countries, and save them as Regular Expressions [3].
3. **Detection of City, State, and Country**. Several names of city, state, and country are collected using MEDLINE and Google search [2].
4. **Organization names**. Organization names are categorized into three levels: Department, School, and University. We collect the words and estimated probabilities of the words for each level as shown in Table II.

TABLE I. AFFILIATIONS IN ARTICLES IN MEDLINE.

| Type | Affiliation | Label Order |
|---|---|---|
| 1 | bioMerieux, Inc. | Org |
| 2 | Faculty of Kinesiology, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4. | Other, Org, City, State, Country, PC |
| 3 | Department of Psychology, University of Houston. | Other, Org |

5. **Other words**. Road name, building number, subdivision name, etc. in affiliations are labeled as Other label. We collects the words for the Other label such as “Avenue”, “Road”, “Street”, etc.

6. **Email**. The following format of Regular Expression [4] is used to recognize emails in affiliations.

```
"([a-zA-Z0-9_\\-\\.]+)@((\\[[0-9]{1,3}\\.[0-9]{1,3}\\.[0-9]{1,3}\\.)|(([a-zA-Z0-9\\-]+\\.)+))([a-zA-Z]{2,4}|[0-9]{1,3})(\\]?)"
```

7. **Hidden Markov Model (HMM)**. HMM [5] is used to label words in affiliations and the Viterbi algorithm [6] is used to finalize the labels of the words in affiliations from the HMM results. The following is the procedure to extract the labels from an affiliation.

Step 1. Separate words from an input affiliation using different seven separators (“,” “;” “:”, “(“, “)”, “[“, “]”).
Step 2. Standardize words (Section IV.1).
Step 3. Assign possibilities of Department, School, and University labels (Section IV.4).
Step 4. Assign possibility of City, State, Country, Postal Code, and Email labels (Section IV.2, IV.3, and IV.6). Assign 1.00 if a word is in the tables or meets its format.
Step 5. Assign possibilities of Other label (Section IV.5).
Step 6. Apply all trained HMMs for the input affiliation and select one HMM ($HMM_{final}$) that has the highest value.
Step 7. Use Viterbi algorithm in $HMM_{final}$ to finalize labels of words in the affiliation.

8. **Statistics Collection**. Several statistics are collected based on the extracted seven labels from each article affiliation such as the number of articles published per country each year, the number of articles per organization each year, the number of articles supported by NIH grants in each country annually. In total, 75 different statistics will be collected.

## V. Web Site Development

A choropleth map and a table are used to display the statistics as shown in Fig 2. The choropleth map uses different shading, coloring, or the placing of symbols within predefined areas to show the statistics in those areas. Google map API, JavaScript, jQuery, AJAX, JSON, are used to develop the map. The left column shows buttons for options to choose among the statistics. Choropleth map is in the center, the top right column shows the color legend for the map, and the bottom right column shows a table that displays the statistics used in the map. The site shows statistics collected from 85,700 articles. As examples, 21,451 articles are published in USA and 7584 articles in China.

## VI. Experimental Results

We use all affiliations in 3,789 MEDLINE citations in the experiment. Of those, 1,767 affiliations are used as the training set and 1,022 affiliations for testing. We develop 30 HMMs from the training set. We group the training set data by the label order and train HMMs for each group. The test results show that 964 are labeled correctly and 58 are labeled incorrectly. The HMMs shows 94.23% accuracy.

We also collect the number of articles published per country each year and the number of articles published per US state each year and visualize them in the Web site.

## VII. Conclusions

This paper proposes a system to show statistics collected from MEDLINE in a Web site using a table and choropleth map. To collect the statistics, an algorithm is developed to automatically extract seven different labels from author affiliations using HMM and statistics. As a future task, we will improve the HMMs by adding more data in the training set and updating word list tables. In addition, we will collect more statistics such as the number of articles related to key words in citations.


## Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.


## References


[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html.
[2] http://www.google.com.
[3] https://msdn.microsoft.com/en-us/library/hs600312(v=vs.110).aspx.
[4] http://regexlib.com.
[5] Chahramani, Z., “An Introduction to Hidden Markov Models and Bayesian Networks”, International Journal of Pattern Recovnition and Artificial Intelligence, 15 (1), pp. 9-42, 2001.
[6] Viterbi, A. J., Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory* , vol. IT-13, April, 1967, pp. 260-269.


TABLE II. Probabilities of Words for University, School, and Department.

| Affiliation Words | Prob. of University | Prob. of School | Prob. of Department |
|---|---|---|---|
| Center, Centre,. | 0.5258 | 0.3196 | 0.1546 |
| Hospital, Hôpital, | 0.7383 | 0.2523 | 0.0093 |
| Institute, Institution, | 0.4779 | 0.4412 | 0.081 |
| University, | 0.9795 | 0.0154 | 0.0051 |

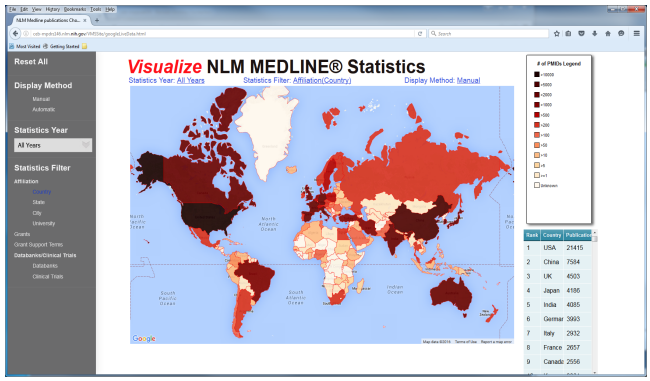


Figure 2. Proposed Web site.