International Journal of Advancements in Technology

Thangaraj and Ganesan, Int J Adv Technol 2016, 7:4
DOI: 10.4172/0976-4860.1000170

Research Article Open Access

Discovery of Drug Brand Names from the Web

Thangaraj M[1*] **and Sivagaminathan P Ganesan**[2]

[1]*Department of Computer Science, Madurai Kamaraj University, Madurai, India*

[2]*Department of Computer Science, Bharathiar University, Coimbatore, India*

[*]**Corresponding author:** Sivagaminathan P Ganesan, Department of Computer Science, Bharathiar University, Coimbatore, India, Tel: +91 8870159676; E-mail: sai.nathan@rocketmail.com

Rec date: July 13, 2016; **Acc date:** July 19, 2016; **Pub date:** August 12, 2016

Abstract

Pharmaceutical drug has been widely prescribed by physicians in providing appropriate chemical composition to the patients for remedial action. Every drug name is being referenced by multiple brand names in the pharmaceutical industry for marketing the same drug with different brand names to elevate a product. This overwhelming competition across the globe for each brand name provoke the doctors, pharmacist, vendors and medical representatives to be familiar with brand names of the same drug as a ready reckoned. In this paper, drug alias retrieval using regular expression has shown significant improvement in Precision and a fair result for Recall, and F-Score.

Keywords: Chemical name; Drug generic name; Brand name/Trade name; Semi-supervised learning; Relevance feedback; Text filtering; Precision; Recall; F-score

Introduction

Internet accessing has become the necessity for a common man, searching for people [1], product, and place by the corporate companies, individuals has increased drastically from daily basis to hours, minutes, and seconds due to the mobility of smart phones, i-phones, and tablets. Some new companies may step in, few companies may move out of the market due to various factors which affects the production of a specific drug which depends on the market demand.

Almost all businesses had gone for sophisticated information systems to integrate sales, inventory, billing and monitoring the nuke and corner of shopping malls, clinics and medical shops. People inquiry minds often changes progressively due to the tremendous usage of electronic gadgets and access to the internet irrespective of age. When internet becomes the global market place for buying and selling, many hospitals buy biomedical instruments, medicines, surgical accessories online.

It becomes a basic need to individuals, to know about purchasing reliable medicines [2] online subject to avoidance of duplicates. In the future, the reach of electronic media and internet are expected to replace conventional newspapers, and magazines. When there is no demand for traditional printing jobs people automatically learn to cope up with electronic media for their daily work. Few years ago, senior citizens felt laborious to operate modern touch screen enabled smart phones. Nowadays, senior citizens are much comfortable to operate the same. Similarly, this technological advancement in the electronic medium will facilitate people to fully depend on electronic applications to access much information instantly. Consequently, this application drug aliasing is experimented using the tool A2E. In web content mining, the purpose of extracting alias names or synonymic keywords of a drug is to use them as a seed to reformulate the query in order to narrow down the search for further expansion and searching. This results in automatic removal of lexical ambiguity yielding potential unknown precise information. Accurate alias identification of a place and object in web is useful in Information Retrieval (IR), Name Disambiguation, Relation Extraction. Also, alias identification of a person in web plays multi-faceted role inclusive of all the three and sentiment analysis. Physicians [3,4], Medical shop vendors and representatives are not in a position to remember all such different brand references since there are large number of drug repository. Moreover, medical professionals, shop vendors need to change their Current Index of Medical Specialities (CIMS), Monthly Index of Medical Specialities (MIMS) directories periodically to know the updated branding names, side effects, diseases for which the drugs are meant, manufacturing companies etc., for each drug as it is prevailing in the current pharmaceutical market. Every drug has at least three names they are Chemical name, Generic name and Brand name.

Chemical name

The chemical name describes the atomic or molecular structure of the drug. This name is normally too complex for a general purpose. Therefore, an official body assigns a generic name to a drug across the globe.

Generic name

The generic or scientific name of a particular drug is the term given to the active ingredient in the medicine that is decided by an expert committee and is understood internationally. A group of brand names that have similar actions often have similar sounding generic names. Therefore, different brand names share the generic name called lexical ambiguity. For instance, Tetracyn, Minocin, Minocycline, Achromycin are brand names belonging to a group of antibiotic tetracycline.

Brand name

The brand name is selected by manufacturer or distributor of the drug. The name is often chosen to be memorable for advertising, or to be easier to say or spell the generic name. For example, Paracetamol is

a generic name. There are several companies that make this with brand names such as Panadol, Calpol. Many drugs have more than one name and therefore, the same drug may be listed more than once in drug directories. Some drugs have too many generics and brand names to itemize on one list. Inclusion of a brand name does not imply recommendation or endorsement. Exclusion of brand names does not imply that it is less effective or less safe than one listed in the CIMS/MIMS drug directories. Few pharmacy websites like eMedExpert, RxList, mims are available to know brand names for a given generic name and vice versa. Nevertheless, there is no specific tool for extracting the brand names of each and every drug that is published on reliable dynamic web.

Combination of drugs

Some drugs or pills contain a combination of medicines. Combined products are marketed and sold with a brand name. Nevertheless, the individual ingredients normally listed in small print on the packet. For example, brand name "Augmentin" is referenced for two generic names such as "Amoxicillin" and "Amoxicillin and Clavulanic acid". Another example, brand name of "Clopidogrel" is "Plavix" and combined product "Aspirin and Clopidogrel" is also known as "Plavix". Therefore, the same brand name given for two different entities in the same domain called referential ambiguity (Figure 1).

Figure 1: Vague Brand names for different Chemicals.

Whenever it is required to list all companies currently manufacturing a specific drug (generic name) for instance 'ciprofloxacin', it is mandatory to collect all brand names of the drug which in turn would expand the query, searches the web for acquiring further knowledge. Hence, unless the user knows complete list of brand names retrieved and ranked in order for a drug, it is impossible to access other details of the drug such as which drug manufacturer is producing a specific brand? What are the pharmaceutical companies using the same generic name? Which branding is fast moving among the customers? etc.

Therefore, this additional information pertaining to a drug can be obtained from the web only by listing the brand names in order. This retrieval is a complex task due to the fact that web is an unstructured electronic medium where it contains heterogeneous content like text, hyperlinks, audio and video files. Besides that, the web designers and developers do not have any common format for preparing text documents.

In earlier research of natural language processing, grammar and lexicons were used for performing text analysis in back volumes, reports and journals, whereas web information extraction [5] uses machine learning and pattern mining techniques to explore the syntactical patterns.

Machine Learning is a subset of computer science that emerged from the study of computational theory in artificial intelligence and study of pattern recognition. Machine learning influences the study and formulation of algorithms where it can learn from and to make predictions on past data. Information extraction, relevance feedback, information filtering, text clustering and text classification are the applications of machine learning in IR.

Drug alias extraction [6] is a subfield of web extraction with an objective to search only for similar, preceding, succeeding, adjacent consecutive set of words in a large collection of medical and pharmaceutical corpus. The term-weighting measure [7] is used to test the accuracy of retrieval in a statistical perspective. In any Information Retrieval task [8], two measures are used in practice to assess the ability of a system to retrieve the relevant and to reject the non-relevant items of a collection called Recall and Precision respectively.

Related Work

The discovery [9-12] of alternate personal name retrieval has been grown on a stage by stage basis and brought to a concrete idea of exploring supervised machine learning approach confining to correct aliases from the web ranked according to the relevant hits. The knowledge discovery process [13,14] initiated the task with the aid of known name-alias pairs held in a training dataset to fetch the similar, synonymic lexical words in a relevant search engine returned brief text characters called web snippets. The architecture of existing alias extraction method is shown in Figure 2.

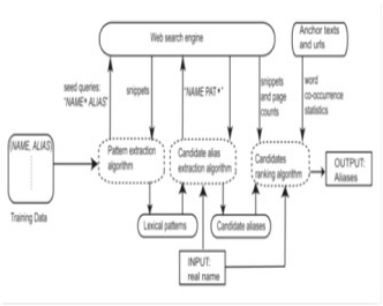

Figure 2: Automatic Alias Extraction using snippets.

Lexical pattern extraction through Google snippets

In the existing method [15], brief text snippet returned by the result of query given by a search engine is the basis for further pattern extraction, candidate alias extraction and ranking. Snippet fetches a couple of lines prior to and few lines successive to the input of search engine. Web snippets contain on an average approximately 250 characters inclusive of spaces and symbols as in Figure 3.

Snippets provide the useful semantic clues that are frequently used in web documents to describe a drug generic name in conjunction with its brand name. It is observed from the snippets that there are some standard text patterns frequently occurs in web documents, which acts as a clue for alias retrieval called formal patterns. Clue like 'also known as' can be used to fetch the brand name as it appears like 'amoxicillin also known as 'amoxil'. Further proven clues are listed in Table 1.

Consequently, Figure 3 shows shallow pattern extraction method where multiple ways in which alias information can be captured expressed as such on the web. Lexico-syntactic patterns have been used in numerous related tasks such as extracting hyponyms [16] and metonyms [17].

Sl no	Formal Patterns
1	[name] nickname *
2	* was born [name]
3	[name] better known as *
4	[name] also known as *
5	[name] alias *
6	* aka [name]
7	[name] aka the *
8	* whose real name is [name]
9	* nee [name]
10	[name] aka *

Table 1: Formal patterns used in existing method.

Penicillins include drugs such as amoxicillin(**also known as Amoxil**), Procaine Penicillin(which is given by injection) and Phenoxymethyl Penicillin (**also known as Phenoxymethyl Penicillin(also known as Phenoxymethyl Penicillin (also Known as PVK)** ....

Figure 3: Snippets returned by Google for the clue 'also known as' in existing method.

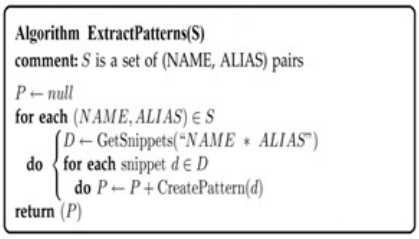

Figure 4: Given a set of (Drug Name, Alias) instances, extract lexical patterns.

Given a set S of (Drug Name, Alias) pairs are considered as input dataset for the Extract Patterns function Figure 4. This function returns a list of lexical patterns that frequently connect names and their aliases in web snippets. For each Name-Alias pair in S, the GetSnippets function downloads snippets from a web search engine for the query "Name * Alias". Then, from each snippet, the Create Pattern function extracts the sequence of words that appear between the name and the alias.

Finally, the real name and alias in the snippet are replaced respectively by two variables (Drug Generic Name) and (Brand Name) to create patterns. The lexical pattern also includes words, symbols and punctuation markers. From the snippet Figure 3, patterns were extracted using the alias marker "(Drug Generic Name) aka *" by the algorithm Figure 4.

To make the retrieval effective the same query is reversed to "* aka (Drug Generic Name)" to extract patterns in which alias precedes the name. Algorithm ignores stop words and stemming words during pattern extraction if present in the snippets. Thus, lexical patterns are extracted for each name.

Extracting candidate brand name aliases

From the lexical patterns available on hand, candidate aliases are extracted for each name as described in procedure Extract Candidates as in Figure 5. Finally, the GetNgrams function extracts continuous sequence of words (n-grams) from the beginning of the part that matches the wildcard operator. It is assumed that generally brand names do not exceed five consecutive words in a document. Hence, first five consecutive words are selected as candidate aliases.

Later, it is trimmed off referring to the training dataset if required to bring meaningful output. For Instance, retrieved Snippet Figure 3 for the query "amoxicillin also known as *", the algorithm Figure 4 extracts 'amoxil' and "phenoxymethyl pencillin" and "PVK" as output candidate aliases from the Figure 5.

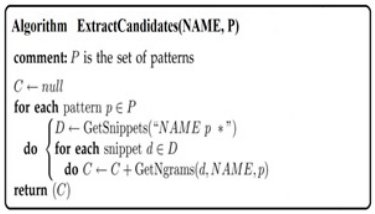

Figure 5: Given a Drug Name and a set of lexical patterns, extract candidate Aliases.

Candidate brand names ranking

As of now there is no formal method of writing text in a document, to maintain uniformity among web documents. It is the ability of the algorithm to discriminate correct versus incorrect output and to disambiguate clearly. From the candidate alias list, it is required to identify which are the most relevant and likelihood to be correct brand names for a given drug.

These candidate aliases were modeled using relevancy in ranking, which one is most relevant to the input query and it is assessed based on lexical frequency, word co-occurrence and page counts on the web. Also, to bring accuracy in ranking this method uses various measures to identify the statistical significance of drug name and its brand name in a retrieval process.

Proposed Work

The method

The outline of proposed method is given in Figure 6, Automatic Alias Extraction being abbreviated as A2E Bot. This A2E bot requires a seed URL to initiate the crawling task. Crawling were done for a period of time, takes a copy of each page that is visited and stored in a binary form to perform pattern matching, candidate alias extraction, and ranking. Prominent feature of proposed method is the use of regular expression to fetch aliases quickly. The tool needs internet for crawling alone, rest of the search process is being done offline and this is the advantage of this implementation.

A2E bot comprises of three major components:

1. Crawler Application
2. Generic Crawler
3. A2E Front End

Choosing the right URL as a seed for each pattern and each drug name, is one of the key aspects to bring efficiency in retrieval process. Only URL which are relevant to the pattern and input drug name must be considered for crawling.

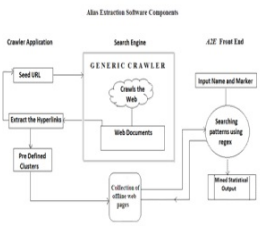

Figure 6: Integration of software components in A2E.

While crawling, most of the web site contains not only pharmaceutical drug content, also might contain medical information, diseases, treatments, advertisements, accounts, payment and other information which are absolutely irrelevant for the intended search query. Thus, the offline storage allows irrelevant pages to some extent which becomes unavoidable to create a feasible web page collection.

The starting seed identifies the subsequent pages to be visited and it traverses recursively till there are no more links on the web. This could be true for few seed URL's but not all the time. Consequently, time frame cannot be estimated to traverse the entire web even for a single seed. For study purpose, crawler is allowed to run for a desired period of time to create a miniature web in the form of database. *A2E* creates ten threads simultaneously to span across the web graph using breadth first traversal. Generic search engine and crawler application continue to crawl which works in connection with each other till the user quits crawler application. In this paper, the maximum number of drug related web pages stored offline are 39,937 with 2500 sports related pages as garbage totally 42,437 textual pages of size 3.15 GB.

Pattern matching technique: The popular pattern matching standard for string parsing and replacement which are used widely in a range a platforms and programming constructs. It is a powerful way to match text with patterns, language independent, written both case and case ignored. Hence, regular expression parses large amount of text documents quickly to identify a specific co-occurring string pattern. The custom designed regular expressions were used in this method for a set of formal patterns in Table 2. A sample Regular Expression of one such pattern is given below

The pattern '(Drug Name) better known as *' can be transformed in the form

\b(drug name)\b.{0,30}?\b(better.{0,3}known.{0,3}as\b

Once the web pages are collected from a crawler it must be classified in to three different clusters similar to the previous alias extraction method like *Sports, Science, and Politics*. But, for simplicity this paper considers only Drug dataset for alias retrieval. Exactly forty seven drug name and various known alias details are stored as training dataset in the tool.

The code initially learns through examples as given in the training dataset inducing supervised machine learning. Over a period of time, it fetches all similar co-occurring patterns wherever pattern matches, hence new aliases are inserted appropriate place of training data with little human intervention. Therefore, this method follows semi-supervised machine learning technique from known to unknown in extracting brand name aliases online. Since web is very dynamic, new brand name may come in and old brand names may move out or it becomes ineffective. Web is the only media, through which any one can get up-to-date information about drugs and medical information. New brand name can be acquired through relevant feedback mechanism [18] and updating. It has been observed that this A2E tool shows relatively higher precision regard to drug name dataset and relatively better F-score with respect to personal name dataset after automating machine learning.

Sl No	Formal Patterns
1	[drug name] aka*
2	[drug name] aka the *
3	[drug name] popularly known as *
4	[drug name] nickname *
5	[drug name] better known as *
6	[drug name] also known as *
7	[drug name] alias *
8	*aka [drug name]
9	[drug name] otherwise known as *

Table 2: Formal patterns used in drug alias retrieval.

Drug data set

With an objective to train the machine, method created a new drug dataset from drug directories and also through official home pages of company sites, and web resources taken from Google.

For each drug name-alias pair stored in dataset it has four alternate spell characters in order to reduce the number of noisy aliases in the output. Since web is very dynamic, any new website may crop up with new alias for a person, in such cases A2E allows user to add as a new row in the training dataset. If the alias is added as a new row, the same

is reflected in alias count and ultimately in recall computation. Otherwise, alias count is not incremented.

Proposed architecture

This proposed architecture Figure 7 needs an offline web archive crawled from the web in order to do further pattern matching, candidate alias extraction, ranking and evaluating performance of proposed method. The entire web page is stored in a non-HTML compressed binary form. An open source.NET code library [19] named Html Agility Pack has been used to parse text documents before searching co-occurring text patterns in the offline storage. Pattern matching and extraction is quickly carried out using regular expression syntax similar to grep command in unix platform. Though it takes time for crawling, usage of regex provoke rapid search and extraction.

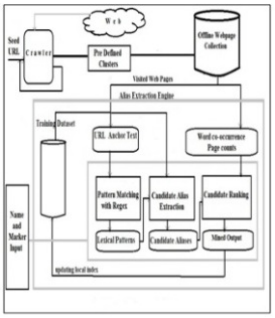

Figure 7: Architecture of A2E bot.

The merit of this architecture is summarized as follows:

- Unique crawler is deployed instead of calling web search engine inside algorithm
- Web archive can be re-usable for another application. Here, Downloaded web pages were held as an attribute of a database table in contrast to search engine returned snippets held in an array
- Candidate alias extraction algorithm need to search within 100 characters succeeding or preceding the regular expression pattern which is much better than reading around 250 characters in snippet array
- For ranking the extracted aliases, lexical-pattern frequency and word co-occurrence count are the same because of usage of regular expression co-occurring syntax "sachin aka *", "sachin better known as *" etc.,
- Automatic removal of stop words and stemming words in IR owing to the usage of regular expression co-occurring syntax. No need to read every line or character to search intended patterns.
- This work is confined only to retrieving textual aliases so that pages containing audio and video files are ignored while crawling.
- Regular expression search is bit powerful than lexical search or keyword search. In existing lexical search, Query "sachin aka *" given within quotes in any search engine results co-occurring wild card character snippets in bold letters with few noise. On the other hand, without quotes returns snippets of both within and without quotes leads to high noisy snippets. When regular expression co-occurrence is used, no extra code needed to eliminate noisy text. However, junk characters cannot be suppressed.

There are two search engines used in this research tool (1) Generic Search Engine (2) Drug alias search Engine. Architecture is designed in such a way that both search engines should work independently. Nevertheless, Drug alias search engine functionality depends on the output of the former one.

Name and search pattern are given as input to the alias extraction engine. Alias extraction engine comprises of modules for catching co-occurring lexical patterns '*(name) aka*' using regular expressions for each pattern. Once the co-occurring patterns are caught, candidate alias extraction is performed using the algorithm Figure 8. For each pattern alias count, noisy alias count, word co-occurrence count, and page counts, precision, recall etc., were computed. Through iterative extraction the training dataset is kept up-to-date in drug aliasing for 47 drug generic names. Using available word co-occurrence statistics and page counts alias names will be ranked in the future for each drug.

Limitations of method

Considering the efficiency, generic crawler used in this method fetches textual content alone by ignoring audio and video contents getting stored in the database. This method ignores few aliases which do not co-occur with drug name and formal patterns. There could be few other patterns left unexplored for this study of alias information retrieval.

This method considers keywords wherever the co-occurring pattern exist in documents followed by accessing the aliases. Since medical and pharmaceutical fields are quite dynamic in nature, most of the brand names were ruled out in today's market or some brand names might not be included in chosen web pages. In such cases, this method considers web as the updated medium, brand names which are available on the chosen pages only are taken for recall computation. For instance, drug *Paracetamol* has around 155 brand names since from the day molecular structure, and drug was formulated. Nevertheless, Google search engine finds hardly two or three brand names which leads to low recall for the same drug.

This method has not restricted number of brand aliases in training dataset for each drug for the sake of getting improved recall. Maximum number of prevailing brand aliases were uploaded in to the training data at par with the CIMS pharmaceutical publications. Hence, the result of low recall is due to the limited entries on the web. Certain brand names appear frequently in web pages rather mentioning every brand name along with formal patterns as expected in this research. For instance, drug-brand pair like metronidazole- flagyl, loperamide-imodium, ciprofloxacin-cipro, cyclosporine-neoral etc., has many entries in crawled samples.

Sample drug-trade name aliases

Drug Scientific Name	Extracted Brand Aliases
Tetracycline	Duramycin, Aureomycin, Minocin, Hostacycline
Paracetamol	Calpol, Panadol, Tm, Tylenol, Crocin, Mol
Gentamycin	Garamycin, Gentak, Gentamicina
Metronidazole	Flagyl, Metrogyl, Monistat
Cyclosporine	Neoral, Atopica

Table 3: Few retrieved trade-name aliases.

Algorithm/Functions

```
Algorithm 2.0 ExtractCandidates(Name,P)
  // P is the set of patterns
  C ←null
  For each pattern p ϵ P do
  Begin
      D ← LoadPage(pattern)
      For each page d ϵ D do
      Begin
          C ←C + GetNgrams(d,Name,P)
          Display mismatched list, candidate aliases, computed results
      End
  End
End
```

Figure 8: Algorithm extracting candidate aliases of length 100 characters.

```
Function 1.1  LoadPage(selected pattern)
  X←dot net implementation of regular expression co-occurrence
  Y←load an offline web page into cache from dataset
  T ← Start a new thread to perform pattern matching
return(Y)
```

Figure 9: Web page loading for a single pattern.

```
Function 1.2 PatternMatch(X)
  //binary form web page is converted in to HTML document form using html
agility pack//
          corr =0,incorr=0
      If( regex(X) exactly matches in webpage)
          corr=corr+1
      elseif (regex(X) partially match in webpage)
          check for ambiguous alias names given in training dataset
      elseif (regex(X) doesnot match in webpage)
          incorr=incorr+1
//pattern co-occurrences are clearly classified in to two sets//
return
```

Figure 10: Matches pattern with actual text in pages.

```
Algorithm 1.0  ExtractPatterns(S)
// S is a set of name-alias pairs//
P ← null
For each (NAME,ALIAS) ϵ S do
begin
      D←LoadPage(pattern)
      For each page d ϵ D do
      Begin
          P ← P+PatternMatch(d)
      End
End
```

Figure 11: Algorithm Extracting Patterns.

Unlike existing method, snippets were not downloaded from the web page to carry out pattern extraction and further candidate alias extraction. Figure 8 accepts drug name and a pattern from user where P is the set of patterns as in Table 2. Regular expression fetches text string of length 100 characters succeeding or preceding formal patterns in each page. Figure 9 provides the code with relevant training data to do pattern matching. Nevertheless, there are few ambiguous aliases which partially match.

In such case, code calls Figure 10 to further classify exactly in to two distinct groups either Relevant or Irrelevant through relevance feedback iterative steps. Similarly, drug name conflict is resolved through relevance feedback for each drug. At any instant of time, tool must ensure updated information about drug aliasing which cannot be found elsewhere in traditional drug catalogs (Figure 11).

Performance Evaluation

Performance analysis

It is obvious that web media designers and writers have the freedom to write a keyword as per their convenience, it leads to huge noisy text information in IR. Discarding irrelevant noisy output from huge collection of documents, resolving ambiguous name aliases, screening the phonetic output were the challenges in this research. Accuracy of any information retrieval is normally evaluated through statistical measures precision and recall, where precision(s) in Figure 12 represents value obtained for a single pattern. Table 4 shows the aggregate F-Score obtained from different dataset and existing method for the common patterns.

$$\text{Precision(s)} = \frac{\text{No of Correct Aliases Retrieved by s}}{\text{Total No of Aliases Retrieved by s}}$$

$$\text{Recall(s)} = \frac{\text{No of Correct Aliases Retrieved by s}}{\text{Total No of Aliases in the Dataset}}$$

$$\text{F-Score} = \frac{2 \times \text{Precision(s)} \times \text{Recall(s)}}{\text{Precision(s)}+\text{Recall(s)}}$$

Figure 12: Statistical measures for alias information retrieval.

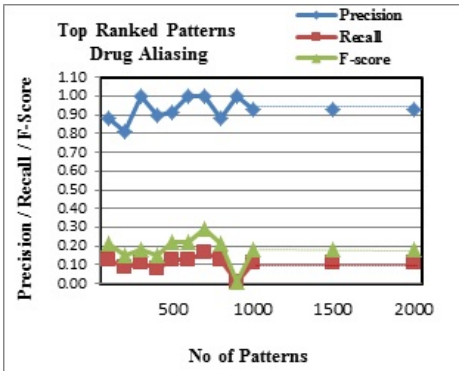

Figure 13: Precision, Recall and F-Score on Drug Dataset.

Unlike personal name alias extraction, drug brand name extraction has comparatively low rejection out of word co-occurrence searching. It could be the fact that, only there are few misspelled or junk brand names on the web as far as drug aliasing is concerned.

Certain names are omitted due to the chemical name or molecular name or user-coined name of the drug. For instance, while extracting brand name aliases, 'Tetracycline otherwise known as 4-

(dimethylamino)', 'Levofloxacin alias cas 100986-85-4 structure molecular formula c18h2', 'Gentamicin aka the Workhorse', 'Paracetamol better known as Blood thinners". Since this research is concerned with extraction of brand names alone, chemical name of drug Tetracycline, molecular formula of drug Levofloxacin and user-defined names Workhorse, Blood thinners were treated as invalid. As a result these names were not updated in to the local index for future references.

Patterns	A2E Drug Name Dataset F-score	A2E Personal Name Dataset F-score	Personal Name Existing Method F-score
aka *	0.21	0.48	0.32
* aka	0.15	0.19	0.34
better known as *	0.18	0.33	0.31
alias *	0.15	0.24	0.29
also known as *	0.22	0.3	0.28
nick name *	0.29	0.36	0.22
aka the *	0.21	0.48	0.19

Table 4: Comparing F-Score with Existing method and *A2E* Datasets.

From the Figure 13, it is obvious that A2E applied on drug dataset has shown significant improvement in precision values compared with A2E applied on personal name dataset. This work proves that drug extraction (A2E) precision outperforms both existing personal name alias extraction and proposed personal name alias extraction (A2E).

Graphical output

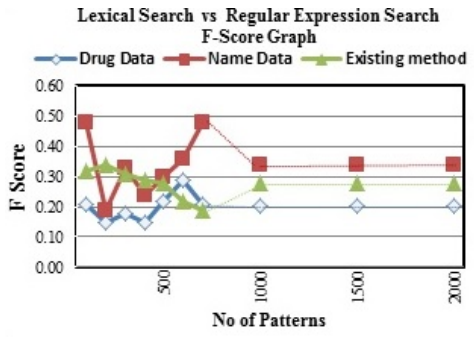

Figure 14: Comparing F Score with existing method and proposed method.

Figure 14 shows the F-Score graph [15,20] for all the three different datasets as in Table 4. A2E tool is experimented with training data on offline web page storage after a long duration of Crawling the web. A seed URL is mandatory for initiating crawling for each pattern-name pair. Subsequently, huge offline storage is obtained for searching desired patterns using regular expressions. Out of 9 patterns as in Table 2, '(drugname) nickname *' yielded maximum value of F-score in this drug aliasing and this pattern is ranked as top most one.

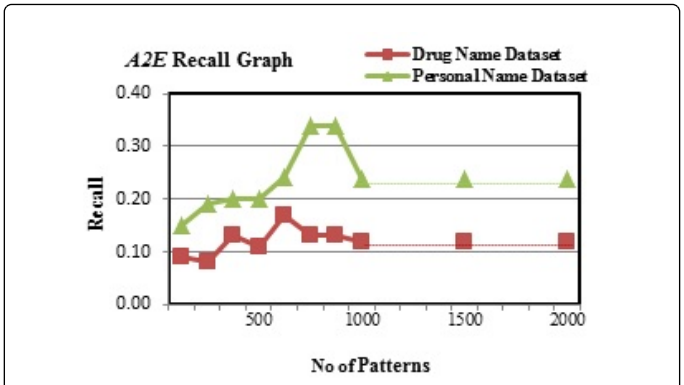

Figure 15: Comparison of Recall on two different datasets.

Figure 15 compares the Recall produced by A2E upon two such datasets: 1) Sports Celebrity Personal Name Dataset 2) Drug Brand Name Dataset. It is evident that Personal Name dataset has given higher recall for a set of common formal patterns. It could be the reason that as of now web is emerging as a media for accessing updated information. Even professionals like doctors, medical representatives have the habit of referring conventional CIMS book to know the updated chemical composition. Nevertheless, there are exceptions among professionals who totally dependent on web for getting up-to-date pharmaceutical industry information across the globe. This low recall has shown the unavailability of brand names of each drug to fulfill the complete pharmacy needs of professionals. In future, not only professionals even a common man will depend on pharmaceutical web forums for collecting authentic information about a drug.

It is evident from the Figure 16, A2E tool outperformed on drug brand name retrieval with few rejections. The average precision of all such chosen pattern is 0.9 out of 1. This high precision is achieved through the novel architecture of alias extraction using quick text processing regular expressions applied on web data.

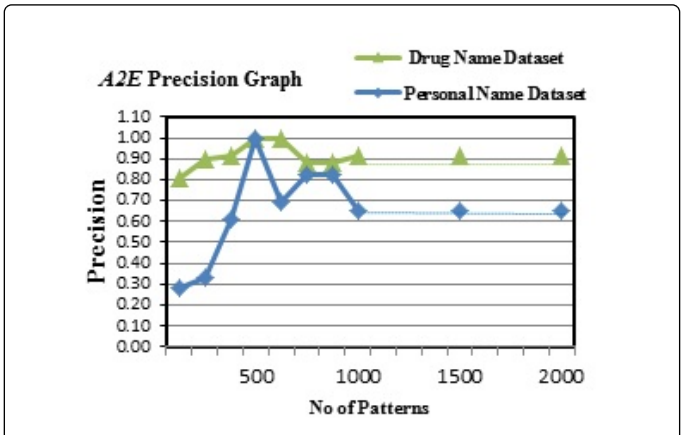

Figure 16: Comparison of Precision on two different datasets.

Conclusion

This architecture has been successfully implemented with four major features they are (a) local archive creation using a generic

crawler (b) relevancy check through proper input seed to the crawler (c) Inducing semi-supervised machine learning and updating local Index (d) Automatic drug name disambiguation in case of referential ambiguity which occurs rarely unlike personal name alias extraction. Although, F-score yielded for drug name retrieval is low, proposed method sports celebrity personal name extraction has shown maximum of 0.48 for two top ranked patterns out of seven common clues.

Proposed method applied on drug dataset is advantageous in bringing out high precision maximum of 1 at two co-ordinates of graph Figure 16 rendering useful in web search task.

Future work

In this paper, only drug domain dataset is used for measuring the retrieval efficiency. Multiple similar domain dataset belonging to such as pharmacy, pharmacology, biomedical, micro-biology may be chosen to cluster each ambiguous name using built-in domain specific keywords. In addition to that, multiple domain web pages can be clustered using methods such as Group Average Agglomerative Clustering, Hierarchical Agglomerative clustering which might be an extension of this work.

For a simple ranking, the most relevant output to the input query can be obtained using co-occurrence statistics, page counts, and cosine similarity measure. To bring accuracy in ranking, Extreme learning machine (ELM) can also be used.

Implementation of fully unsupervised mode of brand name extraction is the open opportunity. Also, ambiguous brand names such as alternate spell characters can be compared using phonetic sound wave forms to carry out classification decision.

References

1. Guha R, Garg A (2004) Disambiguating people in search. Technical Report.
2. Hoffman JM, Proulx SM (2003) Medication errors caused by confusion of drug names. Drug Safety 26: 445-452.
3. Hellerstein JK (1998) The Importance of the physician in the generic versus trade-name prescription decision. The Rand J Econ 29: 108-136.
4. Mott DA, Cline RR (2002) Exploring generic drug use behavior: the role of prescribers and pharmacists in the opportunity for generic drug use and generic substitution. J Med Care 40: 662-674.
5. Chang CH, Mohammed K, Girgis MR, Shaalan KF (2006) A survey of web information extraction systems. IEEE 18: 10.
6. Zhou W, Smalheiser N, Yu C (2006) A Tutorial on information retrieval: basic terms and concepts. Journal of biomedical discovery and collaboration 1: 2.
7. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing and Management 24: 513-523.
8. Salton G, McGill M (1986) Introduction to modern information retrieval. Mc Graw-Hill, New York, USA.
9. Artiles J, Gonzalo J, Verdejo F (2005) A Test bed for people searching strategies in the World Wide Web. International conference on special interest group of information retrieval, Brazil.
10. Bollegala D, Matsuo Y, Ishizuka M (2006) Extracting key-phrases to disambiguate personal names on the web. Lect Notes Comput Sc 3878: 223-234.
11. Hokama T, Kitagawa H (2006) Extracting mnemonic names of people from the web. Lect Notes Comput Sc 4312: 121-130.
12. Bollegala D, Matsuo Y, Ishizuka M (2007) Measuring semantic similarity between words using web search engines. 16th International conference on World Wide Web, USA.
13. Bollegala D, Matsuo Y, Ishizuka M (2007) Identifying people on the web through automatically extracted key phrases.
14. Bollegala D, Honma T, Matsuo Y, Ishizuka M (2008) Mining for personal name aliases on the web. 17th International Conference on World Wide Web, Beijing, China.
15. Bollegala D, Matsuo Y, Ishizuka M (2011) Automatic discovery of personal name aliases from the web. IEEE T Knowl Data En 23: 831-844.
16. Hearst MA, Palo X (1992) Automatic acquisition of hyponyms from large text corpora. 14th International conference on computational linguistics, France.
17. Berland M, Charniak E (1999) Finding parts in very large corpora. 37th Annual meeting of the association for computational linguistics on computational linguistics, USA.
18. Salton G, Buckley C (1990) Improving retrieval performance by relevance feedback. J Am Soc Inf Sci 24: 355-363.
19. Obiwan D (2016) Html agility pack.
20. Thangaraj M, Sivagaminathan PG (2015) A web robot for extracting personal name aliases. Int J Appl Engg Res 34954-34961.