Query Intent Classification using Semi- Supervised Learning

Safallya Dhar
School of Computer Engineering, KIIT University
Bhubaneswar,Odisha,India

Sandeepan Swain
School of Computer Engineering, KIIT University
Bhubaneswar,Odisha,India

B.S.P. Mishra, Ph.D
School of Computer Engineering, KIIT University
Bhubaneswar,Odisha, India

ABSTRACT

The Query Intent classification using semi-supervised learning about ti find a better away to search the web precision that result surfer want to search is 99.8% matched, but due to search engine know what type of query user want to search and logs that are residing in the server of search engine .Which are put in data warehouse of vendor search engine for testing purpose that what type was given. In this paper algorithm is proposed how to increase the precision rate.

General Terms

This paper is classified under artificial intelligence in knowledge acquisition and expert system.

Keywords

Query, intent classification , semi-supervised, tagging, tokens

1. INTRODUCTION

Classification is defined as the categorizing the subjects that are to be pursued or needed to be separate it from the remaining unwanted things. Here in web based classification means need to detect data which the vendor/user demand, this are done by the classifiers which are trained to do classification job. Queries are the important tools of a web surfer to interact with the web but

anyone can't put the queries anywhere as he/she wises it can put only on given search space. Queries are typed in natural language, which are processed by the search engine processor to find the relevant result for the surfer. But the question is how the surfer knows about the indentation of

surfer that what it want to search. So every search based upon user previous log. session it spent on web searching files. So that search engine can know about it intention so next time when user/surfer log search engine can provide better result it provides earlier.

2. REVIEW OF STATISTICAL MACHINE LEARNING

We begin by providing a brief review of basic concepts in statistical machine learning, before presenting a comprehensive overview of the subfield of semi-supervised learning. After explaining the many motivations for wanting to learn with small amounts of labelled data, we discuss some of the most common methods. Machine learning can be categorized into: unsupervised, supervised, semi-supervised method which is followed by different searching engine vendor to make the accuracy rate higher. But now a day's probably many are using semi-supervised learning.

2.2 Basic terminology and notation

An instance X signifies a specific object and is typically represented by dimensional feature vector X = (X_1, . . . , X_D) . Note that boldface X is used to represent the whole instance, and X_d to give the feature of X. A collection of instances {Xi}n i=1 = {x1, . . . , xn} is a training sample and serves as the input to the learning process. To give benefit of X_{id} to represent the i-the instance's characteristics. In most settings, we assume these instances are independently and identically distributed (i.e.) according to an underlying (but unknown to us) distribution P(x). Formally, we write this as: {xi}n i=1 i.e. P(x).Semi-supervised learning lies between unsupervised and supervised learning. Unsupervised learning algorithms work on a training sample with n instances {xi}n$_i$=1. There is null guidance individual instances should be controlled this is the property of unsupervised method. Common unsupervised learning tasks include:

- Clustering: separating the given no. of instances into groups;
- Novelty detection: identifying the instances which are non identical from the rest;
- Dimensionality reduction: For a lower dimensional characteristics vector to represent each instance, To keep the key characteristics of the overall training dataset.

In contrast, supervised methods operates on a training sample consisting of pairs {(x_i, y_i)}ni=1,where yi is parameter on x_i given by nature or some tutor pairs are known as labeled data. Data without labels are known as unlabeled data. Assume the field of instances be X, and the field labels be Y. Assume P(x, y) be an joint probability distribution on data and labels X × Y. Given a training dataset P(x, y), supervised method indulge to solve a function f : X 7! Y in most parameter label, such that f(x) depicts the true label y on future data x, where P(x, y). The two most common types of supervised learning problems are classification and regression. The difference lies in the domain Y. Classification is the supervised learning problem to find a classifier f that can predict one of a set of discrete classes Y. Regression is the problem of learning to predict a continuous value in Y using a learned regression function f. Most of this work will be described in terms of classification, though the methods largely apply to both problem settings. Interactions with search engines reveal three main intents, navigational, informational, and transactional. Giving more précised results depending on such query intents the performance of search engines can be greatly improved. Here the proposed algorithm is described to give more accurate web search results comparing with the existing on.

3. INTUITIVE JUSTIFICATION WORK

The broad area of Query Intent Classification is to classify the data that user wants to search in web as Page & Brin [6] have proposed PageRank algorithm which is said to be model of user behaviour. Brin & Larry assume "erratic behaviour" of surfer to search a particular topic on web and keeps clicking on hypertext and never hitting "back" but eventually gets disgust and gets on random page. The chances that random surfer go to a page is it's PageRank. And "d" damping factor is the probability that each page random surfer will get bored and will request for another random page. Which in thus mislead the system can give higher PageRank. Also there is justification for PageRank is that more number of page cited or directed by other page to a page increases it's PageRank. Google PageRank is developed based on the randomness simply we can say that unsupervised learning, unsupervised is nice technique as here no tagged data is present so no cost and also user van roam around the web without worrying about finite boundary but it lacks the most important part the accuracy in real terms. Here comes the foreplay of Semi-Supervised learning part which takes both tagged and untagged data set. Supervised method works well for annotated data for testing and training. There is null annotated data so can't use the supervised method. This method works one click through possible to train the classifier after starting from smaller sets of queries data. Brin & Larry used semi-supervised method to gain a much larger set of training data set. Trained classifier for searching job on the expanded set.

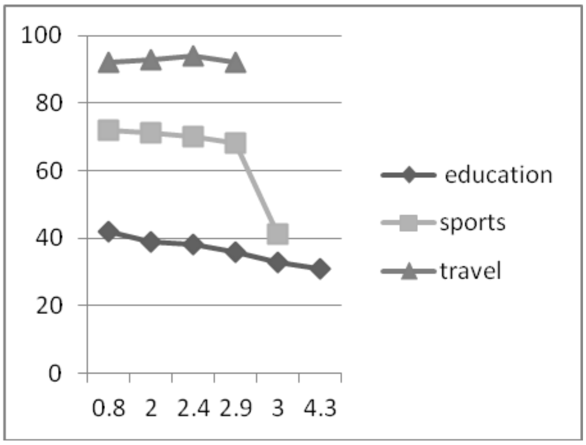

Fig. 1. Depicts the use of relevant word and irrelevant words used in three different categories education, sports, travel.

3.1 Using clicks log as a substitute for annotation

The logs distinguish four types of clicks(a) search results (b)ads (c) spelling suggestion (d) query suggestion. Some prototypical queries of each type are shown in Table 1. The query log contain some terabyte of data every day. A query is typed in same fashion by multi user over a given time period such as click on ads, spelling suggestion clicks.

Table 1. Prototypical Queries of different types

Click Type(Area on Result page)	**Click Type(Area on Result page)**	**Example**
Spelling Suggestion	Type Problem	www.gooogle.in
Ads	Commercial Intent	Flipkart, ebay, snapdeal
Query Suggestion	Suggestible	Phones in 2014 e.g. phones released in 2014
Search Result	Standard Search	The relevant pages

3.2 Query intent

Proposed query intent algorithm based in the user previous log experience on web.

A= User/surfer previous log session files
DB= Database
R= Page repository
i= Pages in repository

Procedure:

```
1. Start processing A in DB.
2. while(A!=θ && R!=θ)
     do
    if (A. anchor = R.i.anchor)
            then return "Match"
       else
            return " Not Match"
3. Gather all the pages in the DB.
4. End.
```

Query intent work of search engine is done by regression training labeled data. This is training is more fruitful is made by semi-supervised learning as now a day's more practices are done through semi-supervised. As semi-supervised understand the sentiment of lexical terms used by surfer so it make perfect conjugation with the accurate page to be searched in web.

3.3 Making search a little more domain Specific

Making the search specification according domain is quite old idea but domain of what it searches in terms of web files, videos, songs, and image. Instead of doing that why we can't create domain like the education, military, cities, religion, and countries. As now also to search a topic which is 20 years old or any scientific research project is as difficult as it was before. We get the files we search for but its accuracy differs. so for advanced search why not to develop a syntactical language or some sort of online tool in common as it will be helpful in finding the exact result it want to search.

3.4 Search using tagging

Token is very useful in developing the search list as words have close associate with the words in approximate letter arrangement and meaning .For e.g. game close associate words are EA sports, gaming zone, action games, adventure games. For quick result search, the search engine also keep some kind of token that might used to speed up the search the user log query ,as search vendor keep the log files of user and the search result of previous session. These token are kept in tables of the search engine during the lexical phase of processing of query it matches the query word with token kept in tables. Token are tagged with the user input word to speed up the search process.

Algorithm: For tagging token with the query word.
Input: Input query is set of j word tokens.
Output: It is result of tagging ($t_i \epsilon T$) attaches to every word token.
$Q=\{w_1[t_1], w_2[t_2],.....w_j[t_j]\}$

Procedure
1. Search for corresponding token to tag the input query word.
2. The tagged word w_i is found then attached it with token t_i for tagging .
3.The corresponding tagged tokens searches for corresponding match in the tables in database of search engine as to create a list of matched token list to which user will find on search engine result list.

3.5 Problems in tagging

Tagging is good for speed up but problems are generated are in more in magnitude so that accuracy is hurt in this tagging process. One of those kind is problem is stemming, stemming is the process of grouping words that share the same root. But now vendor are putting the stemming to user choice. Stemming leads to spamming related issues leads to spamming. Which provides the mingled search result or malicious form of result which user/surfer word might not be interested in.

Input= Query in the form of words in length of size "N" each word containing letters "L"
Output=The list of pages where the query word is found most or probability is more .

Procedure:
```
1. Start
2. Each word $L \epsilon N$
3. While((N-1)!=0)
     do
4. for(i=1;i<=N; i=i+1)
     for(j=1;j<=M: j=j+1)
        if($L_i$ = $L_j$)
                  then
                    return "Matched"
                    Go back to step 3.
        end of for loop
   end of while loop
5. Put all the method letters of word in DB
6.Put value to the page:
     $Page_i=(L_1+L_2+.....L_n)/N$
           if ($Page_i$<0.80)
              return to crawler Table of search engine
7.  if($Page_i(L_i)=Page_j(L_j)$)    //Matching of user query word letter with the token present in the database of the search engine.

then
          for(p=1;p<N-2;p=p+1)
            $Page_i=(L_1+L_2+.................+L_n)/n-p$
             $Page_j=Page_i/2$
               end of for loop
        end of if loop
8.End
```

3.5 Performance Evaluation

User click are checked through and heuristic search is done again to test whether the proposed algorithm hold good to search common query intent whether semi-supervised method put right testing size to check the evaluation of proposed one. Now a days the search are basically one as anyone type one word the meaningful phrase will show so it lessened the user / surfer burden of searching the right sentence and also to check the grammar mistake of the user also auto checked . Because text processing has been so complexes that spider (crawler auto checked the url and the average no. of queries that the user usually search on the web.

Table 2. Table depicting the values of base and proposed algorithm

Values	Base	Proposed
0.4	1	1
0.8	0.9	0.87
0.85	0.886	0.856
0.871	0.472	0.464
0.889	0.299	0.287

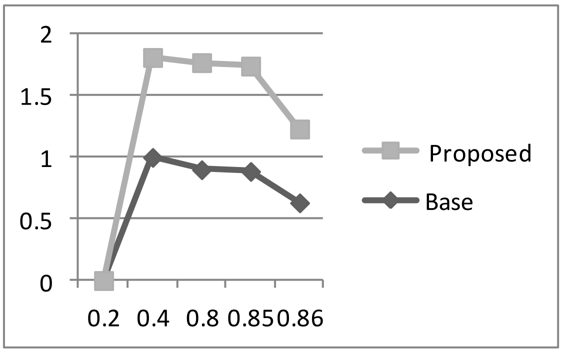

Fig. 2. Showing the comparison between the base and proposed algorithm

The proposed algorithm is compared with baseline both of them perform well when training size data is 100%. Which can be mentioned in figure below:-

Table 3. The training size is done with precision as one of the factor of comparison between baseline and proposed.

Training Size	Precision	
	Baseline	Proposed
100%	85.6	85.4
80%	72.3	79.1
60%	40.8	68.5
40%	28	41.3

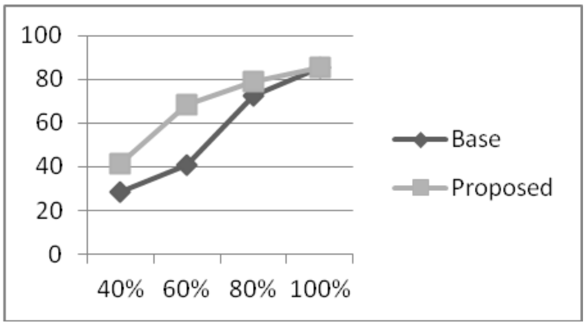

Fig. 3. Precision rate comparison on the basis of training given between proposed and base algorithm

It can be said that proposed algorithm is little better than the baseline algorithm in [2]. But it can't be said that it truly works better than other proposed as it is tested on data size that are based on common query intent such as user click, spell mistake suggestion.

4. CONCLUSIONS AND FUTURE WORK

The performance of search engine can greatly be improved by accurate classifier and data rather than query log. Due to annotated data size all types of data can't be checked though but some common user log, query are put to test the efficiency of the efficiency of search engine. The data of user logs are distorted due to commercial intent of the search engine and the site/url. So in this paper it tries to find out the basic searching technique based upon the existing algorithm. The common user query intent and the accuracy/ precision that the user want result from the web through the search engine. Here the proposed algorithm is compared with the existing one (baseline) [2]. But the future work will be more advanced than this paper as it will be implemented using python which is the language for making a perfect search engine. To test all types of logs/query intent based upon real time user clicks including the commercial clicks through.

5. ACKNOWLEDGEMENT

Special thanks to our guide Dr. B.S.P. Mishra and all esteemed professor of school of computer engineering for guiding us to write this paper. All our friends help and suggestion making this paper reality.

6. REFERENCES

[1] Andrew Brian Goldberg. 2010.New Directions In Semi-Supervised Learning. 24-65.

[2] Emily Pitler, Ken Church . 2009. UsingWord-Sense Disambiguation Methods to Classify Web Queries by Intent. 1428-1436.

[3] Jian Hu, Gang Wang, Fred Lochovsky, Jian-Tao Sun, Zheng Chen.:2009 Understanding User’s Query Intent with Wikipedia. International World Wide Web Conference Committee (IW3C2) 123-142.

[4] S. Brin, and L. Page.1998. The Anatomy of a Large Scale Hypertextual Web Search Engine. Computer Network and ISDN Systems (1998), Vol. 30, Issue 1-7, pp. 107-117.

[5] Dayong Wu, Yu Zhang, Shiqi Zhao.2010 Identification of Web Query Intent Based on Query Text and Web Knowledge. First International Conference on Pervasive Computing, Signal Processing and Applications 128-132,

[6] Shyh-Jier Huang Ching-Lien Huang.1996 Improvement of Classification Accuracy by Using Enhanced Query-Based Learning Neural Network. IEEE398-403.

[7] Ray-I Chang , Pei-Yung Hsiao.1997 Unsupervised Query-Based Learning Of Neural Networks Using Selective-Attention And Self-Regulation. IEEE transactions on neural networks, vol. 8, no. 2, 205-218.

[8] Madhuri A. Potey, Dhanshri A. Patel, P.K. Sinha.2012 A survey of Query Log Processing Techniques and evaluation of Web Query Intent Identification. IEEE 1330-1336.

[9] Arjit De, S.K. Kopparapu. 2012 .A Rule Based Short Query Intent Identification System, TCS-Innovation Lab. IEEE 212-217.

[10] Azin Ashkan, Charles L.A. Clarke, Eugene Agichetin, Qi Guo.2008.Classifing and Characterizing Query Intent. 456-461.

[11] Cristina Gonalez Caro, Ricardo Baeza Yales.2006 A Multifaceted Approach to Query Intent Classification.(2006) 122-124.

[12] David J. Brenes, Daniel Gayo-Avello, Kilian Pérez-González.2009 Survey and evaluation of query intent detection methods,pp557-564.

[13] Raji Sukumar.A, Sarin sukumar A.2010. Key-Word Based Query Recognition In a Speech Corpus By Using Artificial Neural Networks ,pp 212-217.