ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 1, No.19 , January 2016
ISSN : 2322-5157
www.ACSIJ.org

A Simple Study on Search Engine Text Classification for Retails Store

Renien Joseph[1], Samith Sandanayake[2] and Thanuja Perera[3]

[1] Zone24x7 (Private) Limited
460, Nawala Road Koswatte, Sri Lanka
renien.john@gmail.com

[2] Zone24x7 (Private) Limited
460, Nawala Road Koswatte, Sri Lanka
samithdisal@gmail.com

[3] Zone24x7 (Private) Limited
460, Nawala Road Koswatte, Sri Lanka
tm.thanuja.perera@gmail.com

Abstract

It is obvious, the continuing growth of textual content rapidly increasing within the Word Wide Web (WWW). So certainly with the combination of sophisticated text processing and classification techniques it leads to produce high accurate search results. Even though a large body of research has delved into these problems; each has their theories and different approaches according to their data collection. This has been very challenging task continuously and this paper converges solutions, comprehensive comparisons that leads to different approaches. Therefore it will help to implement a robust search engine. The research proves probability text classification models classify documents robustly. But to improve the search result that involves short texts, we should certainly go through a hybrid approach including rules and statistical neural network models. As a pruning components the pre-processing and post-processing modules should adapted. And also due to the dynamic data the process pipeline should be frequently update.

Keywords: Search queries, *Text Classification, Rules, Machine Learning, and Information Retrieval*

1. Introduction

Nowadays the Internet usage is very common. With the rapid growing technologies world is shrinking too small. People do their activities on World Wide Web and they are now getting comfortable with online shopping experience. Hence, it is important the consumers should be provided with concrete online-based search solutions [12][10].

Catering a solution to all kind of consumers is a challenging task. There are consumers use search engine with different behaviorism. In the context of search engine many researchers still have higher involvement to provide a robust solution. Queries can be categorized in to two [9] and they are,

1) Key word/ Short word search queries
2) Long tail search queries

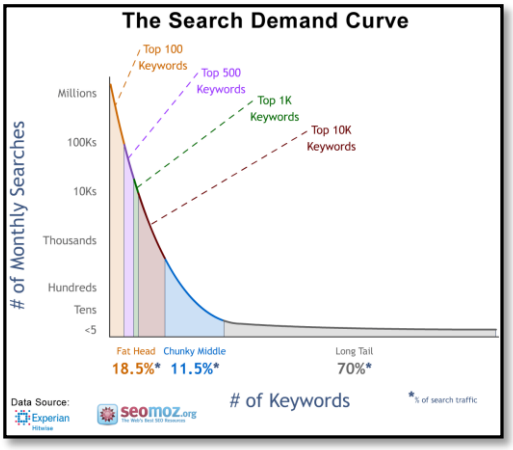

Fig. 1 Search query demand [11]

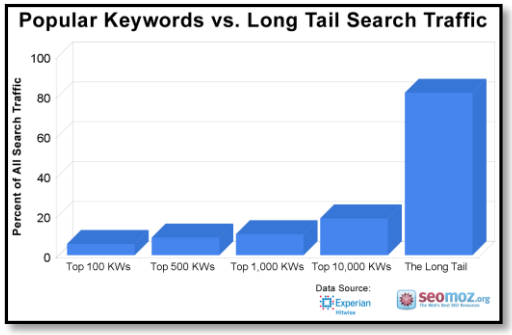

Fig. 2 Benchmarking of search traffic [11]

The graphs (Figure 1,Figure 2) explain the variation of the search queries and its demand. It clearly shows consumers try mostly with short end queries more than long tail queries. Therefore, this research approach can be catered to short and long tail queries appropriately.

During the ionic stages researchers were dealing with handful amount of data and their approach to the search engine was with relational databases [16][21]. But now the evolution has begun in technology and Big Data is extraordinarily growing up in full space [19]. So, with the help of Big Data Analytics [13] this research focuses to provide accurate solution not only to the retail domain but it serve the solution to the various domains precisely.

2. Approach

The data that contains within the retail company will not be enough to implement a robust search engine. Web is increasingly becoming the dominant information seeking method. Every search engine has its root information retrieval (IR) module. IR is all about data retrieval, analysis and emphasizing data as the basic unit (Figure 3). Figure 3 exposes the modules that involves in a search engine component. The indexing component mainly deals with the data mining [2], web mining and social network analysis [4] job that keeps the data collection up to date. The intelligent component tries to understand data collection, user input queries and stay has a decision-making module.

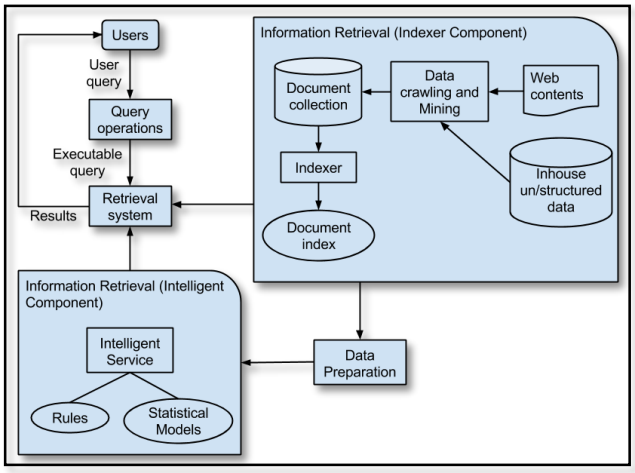

Fig. 3 IR system overall architecture

The main important data process layer is known as Natural Language Processing (NLP) [7]. During this process it will help the engineers to find the information that matching to their domain that helps to boost and improve the search result. In NLP, pre-processing is an inaugural step to processes the query [15]. During this process it helps to clean up and throw away the unwanted words from the query [6]. Figure 4 shows the step-by-step approach for text pre-processing. Due to the pre-processing pipeline, it significantly allows to improve the accuracy in the stage of text classification.

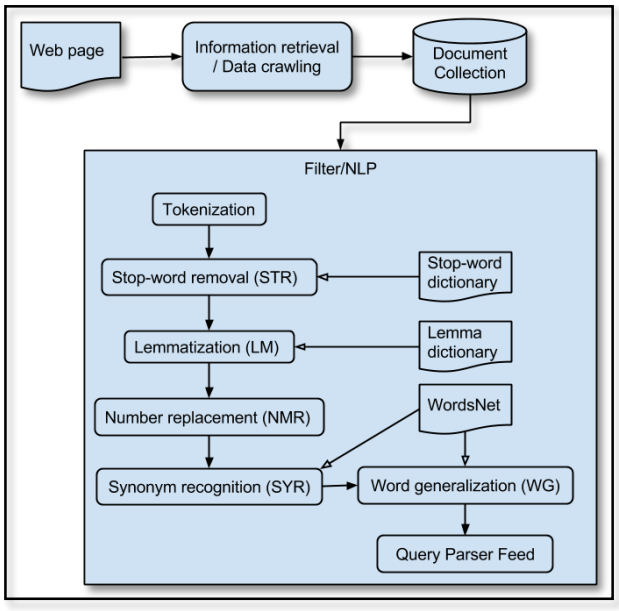

Fig. 4 Text pre-processing schema

In depth, many different components come together according to the users input query. As mention in introduction section the main categories of search queries can be classified in to the following types [15],

- Boolean queries
- Phrase queries
- Proximity queries
- Full document queries
- Natural Language queries

According to the users query usage type, it is important to change the query operational module. In the below sections it covers mostly to tackle the Boolean, Phrase and Natural Language queries.

2.1 Rules Engine

Before the prediction models were invented the developers/researchers were using Rules based search engine. In technology era too still people try to depend on the classical approach to identify the user queries. Rules based engines were implemented focusing into the business domains [1].

The Boolean search queries can be easily handled by rules. The Boolean operators, AND, OR and with negativity words the queries can be constructed. The Boolean operators are filtered during the rules processing stage and certain partial queries are filtered after the Part of Speech tagging (POS) [7]. POS tagging plays an important role in speech and natural language parsing stage. Many researchers attempted to this problem and implemented solution with different approaches [18][17]. Hence, it's very vital to choose the correct approach based on data collection structure and the query patterns also need to be considered during the decision.

2.2 Multinomial Naive Bayes (MNB)

To measure the similarity of two documents, researchers mainly focus on word frequency, which is the most traditional technique. It provides enough word co-occurrences or Shared context for good similarity measures. The Naive Bayes classifier is a simple probabilistic classifier, which is based on Bayes theorem with strong, and naive independence assumptions [14].

This Supervised Learning model falls into Bag-of-Words category and most text classification methods use the bag-of-words representation because of its simplicity for classification purposes [3]. It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. Despite the naive design and oversimplified assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems [3].

In the ambience of retail domain the data type and structure is specific. An abstract data collection is shown in the Table1. After the pre-processing stage the model should identify the feature classes. But due to the sparseness of short text data set list (Table 1), state-of-the-art techniques failed to achieve the desired accuracy [20]

Table 1 Sample data set

Color	Brand	Patterns	Product
Red	Nike	Stripe	Nike Shoes
Yellow	Jennifer Lopez	Stars	Red Jewelry
Green	Hot Wheels	Animal Print	Watches
Purple	American Flyer	Black Squares	Electronics
Black	Nike	-	Jumping Beans
Silver	NORDSTROM	Plane	Sterling Silver Chain
White	Croft & Barrow	Box	Damask Sheet Set
White Lemon	Sleepy Days	Stars	Pajama Set
Khaki	Cuddl Duds	-	Cozy Soft Bed
Black	Wool rich	Circle & Dots	Toys

2.3 Artificial Neural Network

Due to the complex data structure (Table 1) probability theories failed to provide sufficient accurate results. Therefore to improve the result and to deal complex data the focus went towards Artificial Neural Network (ANN). Humans perceive everything as a pattern, whereas for a machine everything is data. So, to think in the direction of human, machines should be trained to understand the data patterns.

Data's are the very important asset for machine and more the data processed the accuracy will gradually will increase or decrease [5]. Most of the prediction problems absolutely focused on pattern matching and recognition. Over the several decades in the filed of pattern recognition, neural network can be regarded as an extension of the many standard techniques. The common approach is to make use of feed-forward network architectures such as perceptron.

2.4 Single Layer Perceptron Model

In its simplest form Single Layer Perceptron is network that classify linearly separable pattern. Various techniques exist for determining the weigh (W) values in single-layer networks. It is widely studied in the 1960's and Widrow and Lehr [22] reviewed and converged for linearly separable pattern.

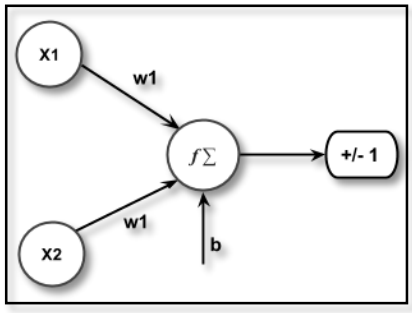

Fig. 5 Perceptron Neural Network

In this basic Perceptron schema (Figure 5) two inputs are the vector elements (x1, x2). The vectors are multiplied with respective weight element (w1, w2) and then summed. This produces a single value that is passed to a threshold function that has only two possible values (1 and -1). 'b' denotes the bias element to prune the model.

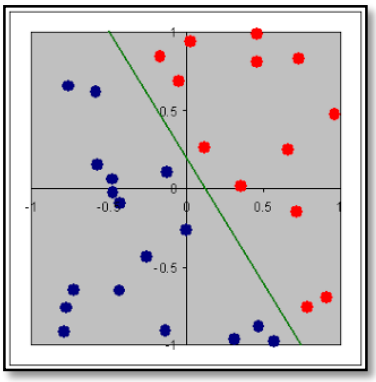

Fig. 6 Linearly separated dataset with Single Layer Perceptron

After Single layer perceptron training; the Figure 6 shows a linearly separated two classes by the green line. The blue dots belong to the first class and the red one belongs to the second. The Figure 6 shows 2 dimension case and by extending to 3 dimension by parse plane that should separate the patterns and dimension that are greater than 3 then it becomes hyper plane that will separate -1 patterns from the 1 patterns.

Therefore according to the retail domain data set (Table 1), it will only produce the correct prediction for data's that converges with AND function (Figure 7) and OR Function (Figure 8).

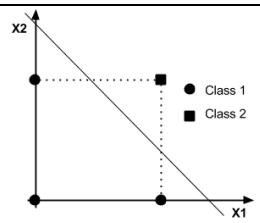

Fig. 7 AND Function

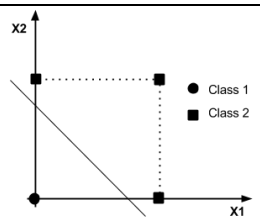

Fig. 8 OR Function

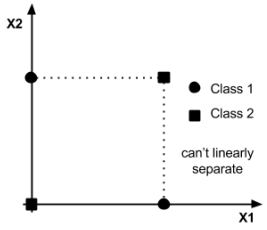

Fig. 9 EX-OR Function

The non-linear separable (Table 1) data that converges with Ex-OR function (Figure 9) cannot be solved with one layer of neural network. So, to improve the classification result need to consider multiple layers and it was proved by Minsky and Papert [5].

2.5 Multi Layer Perceptron Model

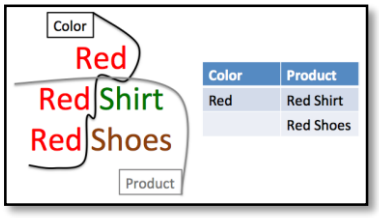

Fig.10 Retail complex data

The above Figure 10 obviously shows the complexity of the data and it's non-linearly spreadable data collection. Therefore MNB and single-layer perceptron models failed to produce a robust prediction.

To allow for more general mapping and to properly classify the data collection need to consider successive transformations corresponding to networks having several layers of adaptive weights. Neurons that are connected to each other's and that send each signal. Commonly neurons can be connected between any neurons and even themselves, but in that cases it gets difficult to train the data set.

Muliti-Layer Perceptron, neurons are arranged into layers (Figure 11). It consists of input layer, one or more hidden layers and an output layer. The signals are passed from one layer to another layer unlit it reaches to the threshold function. The training of the network is done by the highly popular algorithms know as the error back-propagation. This algorithm is based on the error correcting learning rules. Basically there are two passes through the different layers of the networks; forward pass and backward pass.

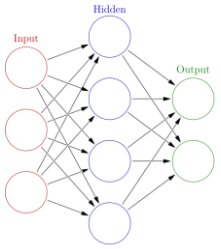

Fig. 11 Neurons mapping

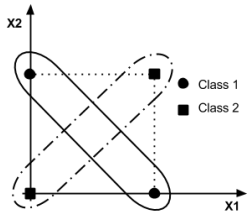

Fig. 12 Separated EX-OR Function

Therefore, with the shortlist of data collection with retail domain Multilayer Perceptron model allowed to classify non-linearly separable data collection (Figure 12).

3. Comparison Of Text Classification Approaches

The test result clearly proves that ANNs are superior over the traditional statistical model when relationship between output and input variables is implicit, complex and nonlinear.

The main strength of each of these algorithms is depended on certain criteria and depends on the context of the domain data. Table 2 presets a concise comparison of all the statistical approaches covered in the above analysis.

Corrêa and Ludermir [8] have done an experimental on ANNs model and the result were extraordinary. The below result (Table 2) expose clearly in general, the MLP Networks distinguished as best classifiers nonlinear data collection [8] and for linear separable data collection the Single Layer perceptron model has much more accurate than Multinomial Navie Base model which is know for probability.

Table 2 Comparison of text classification

Algorithm	Basic techniques	Strengths	Weakness	Highest accuracy reported (%)
Rules	1. Defined rules 2. Defined pattern matching	1. Robust approached for small data set 2. Easy to implement 3. For static data recommended 4. Best for Boolean search queries	1. Dynamic data it fails 2. Difficult to add rule for big data set 3. Can not guarantee	48.6
Navie Bayes	1. Supervised learning 2. Probability based classifier	1. Simple and robust algorithm. □ 2. Independence assumption minimizes computational complexity. 3. Wide applicability 4. Best for document classification	1. Susceptible to Bayesian Poisoning and unrelated video insertion 2. Failed to short text	72.5
Single Layer Perceptron	1. Supervised learning 2. One layer neural network	1. Classify linearly separable data collection 2. Difficult to solve complex problems	1. Non-linear separable data collection can not be classified 2. Lots of training data needed	85.3
Multi Layer Perceptron	1. Supervised learning 2. One or more hidden layer	1. Classify linearly and non-linearly separable data collection 2. Complex problems can be solved 3. Multiple neurons	1. Time consuming to train all the hidden layers	96.7

4. Conclusion

This paper presented a quantitative as well as qualitative comprehensive study of search engine text processing pipeline, text classification techniques and the approaches with the focus of retail domain. This paper includes the accuracy according the dynamic data set that frequently change and the approaches that were taken to make it more accurate. It also assessed the strength and the weakness of the models and the approaches that have taken to improve the search results.

From the research, the observation clearly explicit that significant work has to be done to improve the text class classification models and it obviously going to improve the returning search result which will help the consumers to purchase their needed items within short time period.

References

[1] AGARAM, M. K. & LIU, C. 2011. An Engine-independent Framework for Business Rules Development. Enterprise Distributed Object Computing Conference (EDOC). Helsinki: IEEE.

[2] AGARWAL, S. 2013. Data Mining: Data Mining Concepts and Techniques. Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on. Katra: IEEE.

[3] AGGARWAL, C. C. & ZHAI, C. 2012. A survey of text classification algorithms. In: Mining Text Data. Springer.

[4] ALAVIJEH, A. Z. 2015. The Application of Link Mining in Social Network Analysis. Advances in Computer Science : an International Journal.

[5] BISHOP, C. M. 1996. Neural Networks for Pattern Recognition, Clarendon Press; 1 edition

[6] CESKA, Z. & FOX, C. 2009. The Influence of Text Pre-processing on Plagiarism Detection. International Conference RANLP. Borovets, Bulgaria.

[7] COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K. & KUKSA, P. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12.

[8] CORRÊA, R. F. & LUDERMIR, T. B. 2002. Automatic Text Categorization: Case Study. IEEE Conference Publications. Neural Networks, 2002. SBRN 2002. Proceedings. VII Brazilian Symposium on.

[9] DEMERS, T. 2010. Short Vs. Long Tail: Which Search Queries Perform Best? [Online]. Search Engine Land. Available: http://searchengineland.com/short-vs-long-tail-which-search-queries-perform-best-36762 [Accessed July, 04th 2015].

[10] DREYER, K. 2014. Study: Consumers Demand More Flexibility When Shopping Online.

[11] FISHKIN, R. 2009. Illustrating the Long Tail [Online]. Moz. Available: https://moz.com/blog/illustrating-the-long-tail [Accessed July 04th 2015].

[12] GOOGLE, N. R. S. H. D. C. S. T. L. S. T. W. 2015. New Research Shows How Digital Connects Shoppers to Local Stores – Think with Google [Online]. Available: https://http://www.thinkwithgoogle.com/articles/how-digital-connects-shoppers-to-local-stores.html.

[13] HU, H., WEN, Y., CHUA, T.-S. & LI, X. 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Access, IEEE. IEEE.

[14] KIBRIYA, A. M., FRANK, E., PFAHRINGER, B. & HOLMES, G. 2005. Multinomial Naive Bayes for Text Categorization Revisited. AI'04 Proceedings of the 17th Australian joint conference on Advances in Artificial Intelligence, 3339.

[15] LIU, B. 2011. Web Data Mining, SIGKDD Explorations, springer

[16] LUO, Y., WANG, W. & LIN, X. 2008. SPARK: A Keyword Search Engine on Relational Databases. Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE.

[17] MAHAR, J. A. & MEMON, G. Q. 2010. Rule Based Part of Speech Tagging of Sindhi Language. Signal Acquisition and Processing, 2010. ICSAP '10. International Conference Bangalore.

[18] P.J, A., MOHAN, S. P. & K.P., S. 2010. SVM Based Part of Speech Tagger for Malayalam. Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference Kochi, Kerala.

[19] PARK, H. 2014. Bigger, Better, Faster, Stronger: The Future of Big Data [Online]. cmswire. Available: http://www.cmswire.com/cms/big-data/bigger-better-faster-stronger-the-future-of-big-data-027026.php 06 March 2015].

[20] QUAN, X., LIU, G., LU, Z., NI, X. & WENYIN, L. 2010. Short text similarity based on probabilistic topics. Knowledge and Information Systems, 25.

[21] WANG, W., LIN, X. & LUO, Y. 2007. Keyword Search on Relational Databases. Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference IEEE.

[22] WIDROW, B. & LEHR, M. A. 1990. 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. IEEE.

First Author Graduated from University of Westminster and holds BEng (Hons) in Software Engineering, Computer Science. Currently works as a Senior Software Engineer at Zone24x7 (Private) Limited. Have a quite a good experience in Research and Development side. An active developer with very broad knowledge and keen interests towards Big Data, NLP, 3D visualization, Computer vision technologies, modern web and mobile technologies.

Second Author Bachelor in Computing Science degree holder from the Staffordshire University, UK and currently a Senior Software Engineer at Zone24x7 (Private) Limited, who is passionate in data science and actively researching and developing in projects related to NLP distributed systems.

Third Author A Software Engineer currently works at Zone24x7 (Private) Limited and obtained B.Sc. (Hons) special degree in Computer Science from University of Sri Jayewardenepura. Interested key areas are NLP, Big data and web technologies.