Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 2 No. 2, December 2017, pp. 73 – 79

Naive Bayes Classifier Based Geographic Information System for University Search Information

Junta Zeniarja*[1], Ardytha Luthfiartha[2], Catur Supriyanto[3]
Faculty of Computer Science, Universitas Dian Nuswantoro
E-mail: junta@dsn.dinus.ac.id[1], ardytha.luthfiartha@dsn.dinus.ac.id[2], catur.supriyanto@dsn.dinus.ac.id[3]*
**Corresponding author*

Abstract - *Information about the geographical location of universities is necessary for graduates of Senior High School who want to continue their education to a university. Most of the graduate students do not know the location of the universities since the geographical location of Google Maps is less clear and less precise. Therefore, the application of Geographic Information Systems (GIS) based on Information Retrieval (IR) is expected to facilitate the graduate students to know the exact location of the university. In this paper, IR-based GIS application is developed by using web programming. The web is used as a search engine when someone wants to find a college. The application shows the map and information of the college in the area according to the query of the user. Naive Bayes algorithm is used to classify the user query and locate the query on the map. Based on our prototype, the application is promising to be implemented for the student.*

Keywords – *Naive Bayes, Geographic Information System, University Location Information*

1. INTRODUCTION

The rapid development of the internet brings many changes to human life. Almost every human activity uses the internet. According to market research institute e-Marketer[1], the number of internet users in Indonesia reached 83.7 million people in 2014, making Indonesia ranked the 6th largest in the world. Overall, the number of internet users worldwide is projected to reach 3 billion people by 2015. Three years later, by 2018, an estimated 3.6 billion people on Earth will access the internet at least once every month.

Based on statistical data internet users above, then the internet has a very big role in human life and has become a part of life of humans in the world. One of the benefits of using the internet in the field of science is to access and search for various kinds of information. Most people search for information about something via the internet with the help of search engines like Google. Especially Senior High School students who want to find information about universities either public or private universities that they want to search through the internet. Based on data from the Directorate General of Higher Education in Indonesia, in 2012 there are 3150 universities both public and private universities and 15,830 courses. Currently there are 4,438 universities consisting of 1,105 academies, 241 polytechnics, 2,421 high schools, 130 institutes, and 541 universities. The facts show that the scope of research universities in Indonesia is wide enough. The number of college distribution is also a potential and a challenge for the government to provide a fair and equitable education for the people of Indonesia.

[1] https://www.emarketer.com/Article/Internet-Hit-3-Billion-Users-2015/1011602

Unfortunately, there is still a lot of media that specifically provides information about public or private universities in Indonesia. So that high school students have difficulty in finding the complete and up to date information about the universities in Indonesia. Students also need a kind of internet map application that can pinpoint the detailed public and private universities location they are going to visually, making it easier for them to locate the location.

In terms of search, closely related to the concept of Information Retrieval (IR). IR is the study of methods for rediscovering stored information from relevant sources or collections of sources of information. In data search, some types of data can be found such as text, tables, images, video, and audio. The purpose of IR is to satisfy user information to retrieve relevant documents or reduce unrelated search documents.

Supervised machine learning approach is able to build geographical information retrieval for university search information. There are some supervised method, such as Naive Bayes (NB), K-Nearest Neighbor (KNN), Artificial Neural Network (NN), Support Vector Machine SVM), and Decision Tree (DT). This paper proposes NB to classify the user query since NB has good performance in term of speed and the ease of huge data access [1] [2]. In the similar work, NB has been used to classify some documents into their categories [3]. NB is also success in many application, such as tweet sentiment analysis [4] [5], spam detection [6], and opinion mining [7]

Geographic Information System Applications (GIS) can be built with IR concepts through NB classifier, which by making the category text for maps allows for easy searching of college addresses. For example by categorizing the college based on geographical location, type of college and other categorization. With the categorization of text in IR using NB classifier applied to GIS application to search university information, it is expected that students who wish to continue their education to university will be facilitated in searching the location of information and address only with the minimum information they know.

2. RESEARCH METHOD

2.1. Information Retrieval (IR)

Information Retrieval aims to retrieve the documents, search the documents, search the metadata describing documents, or search in databases, either stand-alone database relations or hypertext databases contained in the internet, for text, sound, images, or data. In principle, information storage and rediscovery of information are simple. Suppose there is a place to store documents and a person (user) formulate a question (request or query) the answer is a set of documents containing the necessary information expressed through user questions. Users may obtain the documents they need by reading all the documents in the storage, storing relevant documents and discarding other documents.

This is a perfect retrieval, but this solution is not practical. Because the user does not have the time or does not want to spend his time reading the entire document collection, despite the fact that the user is physically impossible to do so. Therefore, an information retrieval system is required to help users find the documents they need. It is a system by applying a functionality similar to the ranking of a set of items (usually textual documents) in response to user requests.

2.1. Geography Information System (GIS)

GIS is a combination of three main elements of the system, information and geographic. GIS is a system that emphasizes the element of geographic information, where the geographical

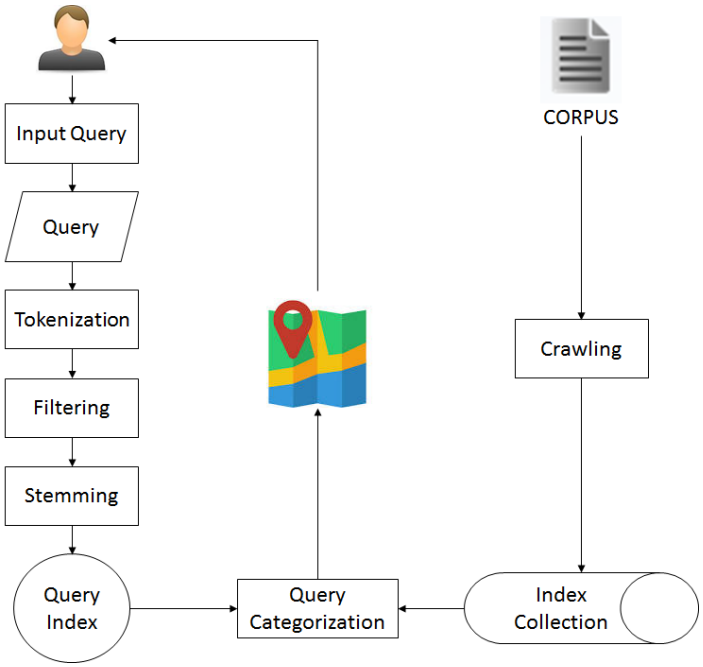

Figure 1. Flow of Our System

information contains the understanding of information about the places on the surface of the earth, the knowledge of the location of an object on the surface of the earth, and information about the attributes contained in the surface of the earth whose position is known. Since the late 1990s GIS began to be integrated with spatial database systems in what seems to be a perfect idea. As a result, GIS is currently focused on spatial data collection, editing, analysis and visualization, while spatial databases deal with data storage, querying, indexing, optimization, and integrity.

In addition, GIS as a system designed to capture, store, manipulate, analyze, manage, and present all types of geo-referenced data, has been widely used as a spatial decision support tool. However, the traditional one-user GIS interaction mode limits the complexity of spatial problems solved and for problem-solving efficiency. In fact, many spatial and decision-making issues involving geographic information require multiple users to work together to process and analyze geographic data. In the process of decision-making and spatial problem solving, there is a tendency to integrate real-time and collaboration, which has become one of the important areas of research and development in GIS theory and applications.

2.2. Text Classification

Text classification is one of the topics in Information Retrieval as well as text mining which has gained significant popularity over the past decade. One of the main reasons is the increasing number of digital documents and the need to access content in a more flexible way. Additionally, text classification is also referred to as text categorization or document classification. The current approach to classify text is a machine learning paradigm using a set of previously categorized documents to automatically build a categorizer by learning from the data. As part of this process, each text document is represented by a feature vector, thus ignoring the

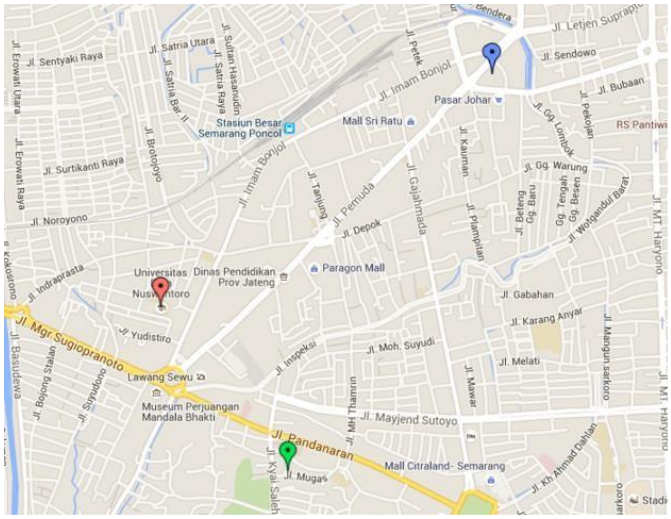

Figure 2. Query Result on Map Location

order of words and other grammatical issues, as this representation is capable of storing enough useful information for classification tasks.

2.3. Naive Bayes Classifier

Naive Bayes is a supervised learning algorithm in the machine learning [8]. Specially, NB is being used for classification. It needs training process to classify the testing data.

$$P(C_i|d) = P(C_i) \prod_{k=1}^{n} P(w_k|C_i) \qquad (1)$$

$$P(w_k|C_i) = \frac{n_k+1}{n+m} \qquad (2)$$

Where $P(C_i|d)$ is the probability of a testing document d or a query in category C_i, w_k is the w-th word in a testing document or query, n_k is total number of word w_k in C_i, n is total number of word in C_i, and m is total number of vocabulary.

3. RESULT AND DISCUSSION

In the Figure 1 on research design of GIS based information retrieval Model, explain some stages in preprocessing from text search with information retrieval concept. The initial process starts from the user who will perform the input query in the form of information to be searched and the user will get output in the form of maps. The operation of preprocessing with information retrieval concepts includes tokenizing, filtering, and stemming. Where when the user input query, then the first operation is tokenizing is the process of solving a word from a sentence. After going through tokenizing, the selected word will be in the filtering operation, irrelevant words will be removed. Then, the selected word will be applied to the stemming process or the basic word search of the selected word. Words that have passed all three operations will converge into an index or query index.

Table 1. Sample Documents

Category	Doc
Semarang Tengah	Universitas Dian Nuswantoro udinus adalah sebuah perguruan tinggi lokasi Semarang Tengah didirikan tahun 1996, akreditasi B, merupakan universitas swasta terbaik di indonesia.
Semarang Timur	Universitas Diponegoro undip adalah pergurukan tinggi Negeri lokasi Tembalang Semarang Timur akreditasi A didirikan tahun 1994 merupakan universitas terbaik di indonesia.
Semarang Selatan	Universitas Negri Semarang unnes merupakan perguruan tinggi negri lokasi semarang selatan, ngaliyan, gunung pati akreditasi B terbaik di indonesia didirikan tahun 1995.

The output of this application is a geographical information system where the user input the query. Then, the query is classified based on the training documents (See in Table 1). The category of the query will be displayed in the form of maps.

The example below shows the classification of the query “Universitas Gunung Pati”:

$$P(Universitas|Semarang\ Tengah) = \frac{2+1}{18+28} = 0.0065$$
$$P(Gunung|Semarang\ Tengah) = \frac{0+1}{18+28} = 0.0217$$
$$P(Pati|Semarang\ Tengah) = \frac{0+1}{18+28} = 0.0217$$

$$P(Semarang\ Tengah) = 0.333$$

$$P(Semarang\ Tengah|Query) = 1.027369113175\text{E} - 5$$

$$P(Universitas|Semarang\ Timur) = \frac{1+1}{9+28} = 0.054$$
$$P(Gunung|Semarang\ Timur) = \frac{0+1}{9+28} = 0.027$$
$$P(Pati|Semarang\ Timur) = \frac{0+1}{9+28} = 0.027$$

$$P(Semarang\ Timur) = 0.333$$

$$P(Semarang\ Timur|Query) = 1.3161444863417\text{E} - 5$$

$$P(Universitas|Semarang\ Selatan) = \frac{0+1}{11+28} = 0.0256$$
$$P(Gunung|Semarang\ Selatan) = \frac{1+1}{11+28} = 0.0513$$
$$P(Pati|Semarang\ Selatan) = \frac{1+1}{11+28} = 0.0513$$

$$P(Semarang\ Selatan) = 0.333$$

$$P(Semarang\ Selatan|Query) = 2.2477340031581\text{E} - 5$$

Based on the example above, the query is classified into Semarang Selatan.

Data source is an important thing in developing a university search application. In conducting data analysis, the authors found several areas categorized into 4 categories, namely *Semarang Tengah*, *Semarang Barat*, *Semarang Timur*, and *Semarang Selatan*. The application aims to make easier for users to search universities in Semarang area. Using the Naive Bayes text categorization method. In this case the text category determines the area in the area of Semarang.

From this research, the system has 2 stages, the first stage involves learning or training that is the classification stage of the documents of several universities in Semarang already in the know category. The testing stage is performed by classifying documents or query into several categories. The result based on the query will be show in Figure 2.

4. CONCLUSION

The use of NB algorithm for query with text category where to measure the closest distance between queries with text category based on input query from user, with NB algorithm can be determined query belongs to which text category. Using the Google Maps API will display the map based on the similarity between the query and the text category. Then done categories of maps based on text categories that have been made. Query the text category of maps as Outputs that come from the Google Maps API category along with the descriptions of the high gallery.

5. FUTURE WORK

In further research, we will propose to use other algorithms such as K-Nearest Neighbor and Weighted Naive Bayes to compare with previous research. It is hoped that the research will get better results than ever before. In addition, we will develop this system with a better platform, combined with mobile applications, making it easier for users to be able to use it at any time.

REFERENCES

[1] Z.-L. Xiang, X.-R. Yu, A. W. M. Hui and D.-K. Kang, "Novel Naive Bayes based on Attribute Weighting in Kernel Density Estimation," in *Joint 7th International Conference on Soft Computing and Intelligent Systems, SCIS 2014 and 15th International Symposium on Advanced Intelligent Systems, ISIS*, 2014.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.,* vol. 14, p. 1–37, 2007.

[3] Q. Jiang, W. Wang, X. Han, S. Zhang, X. Wang and C. Wang, "Deep Feature Weighting In Naive Bayes For Chinese Text Classification," in *Proceedings of CCIS2016*, 2016.

[4] A. Goel, J. Gautam and S. Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes," in *2nd International Conference on Next Generation Computing Technologies (NGCT-2016)*, 2016.

[5] M. Mertiya and A. Singh, "Combining Naive Bayes and Adjective Analysis for Sentiment Detection on Twitter," in *International Conference on Inventive Computation Technologies (ICICT)*, 2016.

[6] D. D. Arifin, Shaufiah and M. A. Bijaksana, "Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier," in *IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2016.

[7] K. M. A. Hasan, M. S. Sabuj and Z. Afrin, "Opinion Mining using Naïve Bayes," in *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2015.

[8] P. Yildirim and D. Birant, "Naive Bayes Classifier for Continuous Variables using Novel Method (NBC4D) and Distributions," in *IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings* , 2014 .