AUTOMATIC CHART INTERPRETATION WITH AN APPLICATION IN HEALTH SURVEILLANCE

Wahyu Pratomo[1], Ari Moesriami Barmawi[2], Hertog Nugroho .phd[3]

[1]Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak
Interpretasi grafik memiliki peranan yang sangat penting di berbagai aspek, misalnya pada suatu penelitian, pembacaan dan interpretasi terhadap suatu grafik akan sangat berpengaruh, baik untuk pengambilan komentar maupun tanggapan terhadap suatu hasil penelitian, hingga penentuan arah improvement yang harus dilakukan. Di bidang sosial pun, pembacaan grafik memiliki peranan penting. Seringkali suatu keputusan diambil berdasarkan hasil penginterpretasian suatu grafik. Banyak kesalahan terjadi akibat kesalahan dalam menginterpretasikan suatu grafik, sehingga pengambilan keputusan menjadi jauh dari kebenaran yang diharapkan karena informasi yang didapat dari grafik menjadi kurang tepat ataupun kurang lengkap akibat grafik tersebut dibaca dengan cara yang salah. Berdasarkan masalah akan pentingnya pembacaan suatu grafik, serta belum adanya suatu sistem pembacaan grafik yang secara otomatis mampu membaca grafik yang diberikan, maka pada riset kali ini, akan dirancang suatu metode untuk menginterpretasikan grafik secara otomatis. Perbaikan yang dilakukan dari metode sebelumnya terdapat pada inputnya yang berupa citra digital dan perbaikan juga dilakukan pada knowledge base yang dibangun berdasarkan informasi yang terdapat pada judul grafik, legend, serta judul pada sumbu x dan sumbu y, disamping juga masih tetap menggunakan corpus sebagai bantuan penyusunan kalimat.

Kata Kunci : Acquisition, Chart, Reading ,Value

Abstract
Chart interpretation has an important role in many aspects. For example, in research, reading and interpreting a chart will give significant effects, not only to take a comment and give a response, but also important to take the performance improvement. Chart interpretation also has an important role in the healthcare. Many decisions, judgments, or conclusions have been taken based on chart interpretation. Many mistakes were made because of misunderstanding about chart content, such that the conclusion which has been taken become incorrect or far from the truth. This condition happened because of much information was resulted from various incomplete or inaccurate information obtained to draw a conclusion or decision. This thesis proposes method for interpreting chart image automatically. The improvement from the previous methods are that the input can be in a form of image and the information used for the knowledge uses information from the chart legends, title and values extraction, beside the corpus.

Keywords : Acquisition, Chart, Reading ,Value

CHAPTER 1: INTRODUCTION

1.1 Rationale

Chart interpretation has an important role in many aspects. For example, in a research, reading and interpreting a chart will give significant impact, not only to take a comment or giving a response, but also important for taking the critical decision. Chart interpretation also has an important role in the health care. Many decisions, judgments, or conclusions have taken based on any chart interpretation.

However, nowadays chart is still a problem because the charts are normally interpreted manually. In rural areas, often there are no professional available to interpret charts. The non-expert person, usually those who have non-engineering education background, often interpret charts incorrectly. This misinterpretation can lead to a bigger problem while the conclusion is far from the fact because of many information which is gained to support a decision is incomplete or inaccurate [1].

Considering the important role of chart interpretation and also from the exploration so far that none of automatic chart interpretation based on Indonesian language, then this thesis entitled *Automatic Chart Interpretation with an Application in Health Surveillance* conducted. This thesis will discuss on an automatic chart interpretation system, designed and developed in Indonesian language, and displayed in most natural language so that any people in Indonesia can interpret the result easily. According to the title of this thesis, the chart is limited to the health related chart. The sample of the chart is illustrated in Figure 1.

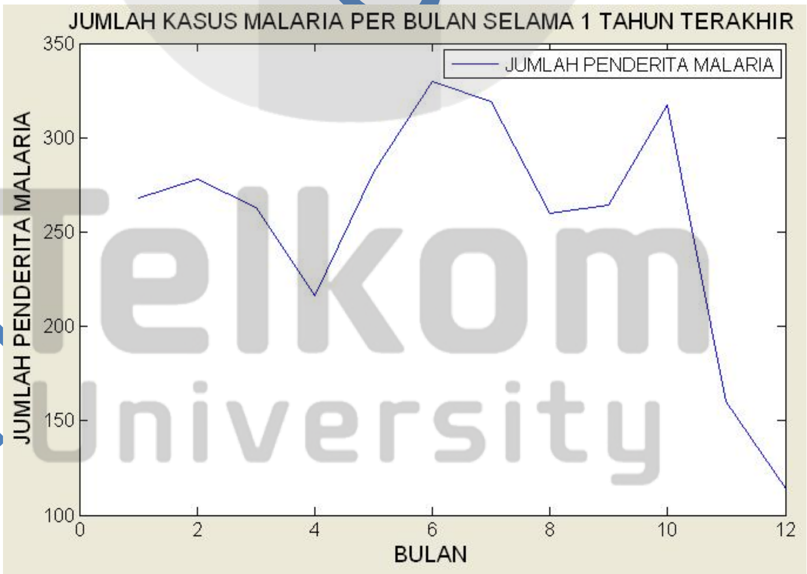

Figure 1 - Chart Example

1.2 The Problem

Chart interpretation nowadays is still a problem, especially to a people with non-engineering education background. Interpretation of a chart is often to be very limited when a given chart is actually full of meaning. This condition is very unfortunate because the reader is unable to understand what the content is actually conveyed by the chart.Incorrect interpretation on the chart will give the wrong perception of the given data. This mistake certainly has significant impact on the decision to be taken based on that chart interpretation.

The problems also come from the Natural Language Processing methods where the existing NLG systems used very large size of the corpus as a language direction. In addition, there are no grammatical rules used in existing NLG systems. The details of this thesis problem are as follows:

1. How can the chart image be read automatically?
2. How can a NLG system with a small corpus be built?

1.3 Theoretical Framework

Given an image such as depicted in Figure 1, then the information to be extracted from the image would be represented by collection of pixels. Certain information can be extracted by examining the layout and structure of the pixels. Symbols are represented by character-like structure while the curve is represented by un-connected layout of pixels. The extraction tools should pay attention on this specific structure and layout and hence appropriate tools should be carefully selected. Much of the algorithms developed for this kind of task fall into the area of Image Processing. To extract character-like symbols, OCR can be used while in extracting curve, certain pixel detection tools should be developed.

If the symbols and values information representing the curve have been collected, a report can be formulated from them. Equipped with appropriate specialty, human being can perform such task without any difficulty. However, the same task would be very difficult to be done by computer, especially when the report has the same quality as those written by a human being. To generate such report, some aspects must be considered such as the grammar, dialect, criticality of the information, etc. Works on these problems fall into the area of Natural Language Processing. To generate a title, certain type of words are selected from the available information, while generating curve analysis result, some processes such as chart value extraction, pattern detection, etc should be done first.

1.4 Conceptual Framework/Paradigm

The main tools used in this thesis consist of Image Processing and Natural Language Processing parts. The image processing part is used to extract the features from the chart image. The features extracted from the chart include chart title, chart legend, axis title, axis number, and the pairs of (*x, y*) number representing by the chart line. The text features such as title, legend, axis number, and axis title, are extracted by Optical Character Recognition (OCR) process. The next process is extracting the line coordinate after segmenting the ROI of the chart line. The challenge is that the (0, 0) coordinate in the Cartesian plane is on the bottom left corner while the (0, 0) position in the image coordinate is on the top left corner, so the line coordinate must be converted. After this process, all the texts from the chart is extracted and stored in a String variable and the pair (*x, y*) features are stored in an integer variable.

The next process is Natural Language Processing. The input for this process is the chart feature variable. The process is done by developing the sentence based on the seed words extracted from chart image. A Seed word is selected by the user to make sure that the system is only developing the sentence needed by the user. The generation is corpus based and the calculation is based on Bi-Gram probability. The result from this process, including the result from this overall system, is the sentence that describes the chart automatically.

1.5 Hypothesis

This thesis proposed a method for interpreting chart image automatically using the combination between image processing and Natural Language Processing (NLP). On image processing side, the OCR method proposed in [2][3][4] can be used to extract and recognize the character contained on the image. Then the OCR is used to extract the text from the chart such as chart title, axis title, axis number, and legend. Chart value is extracted using the calculation of chart line pixel according to the chart scale pixel position [5].

In [6], [7], [8], there are discussions about method to perform an automatic Natural Language Generation System. Based on their researches, the automatic chart interpretation system is realized using Natural Language Processing. The combination using Bi-Gram sentence and grammar rule methods results the more appropriate sentences because the sentences not only constructed using the sequence of sentences, but also calculates the grammar rule to make sure the grammar suitability.

The improvement from the previous methods are that the input can be in a form of image and the information which is used for the knowledge uses not only the corpus, but also information from the chart legends, title and values extraction, in addition to the corpus.

1.6 Assumption

According to the limitation of this thesis, then it is assumed that the chart image is not tilted and its quality is not degraded, for example, by noise. To make sure that the process can run properly, it is assumed that the original data is available and based on that data, the chart can be generated by the Matlab software. Information in the chart image is composed of chart title, legend, axis title, axis number, and chart line. It is assumed that the difference between axis, background, and chart line is clean and clear.

1.7 Scope and Delimitation

The chart is limited to two dimension (2D) chart, which is in the form of time series Chart (no two pair of (*x, y*) value which is exactly the same). Another limitation is that there is only one curve on the chart.

1.8 Importance of the Study

This study will contribute to automatic chart interpretation systems, which can help one in evaluating image-based data. The developed tool can also be regarded as an expert system or decision support system because it is able to produce recommendation based on its interpretation on the image data.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The conclusion from this thesis research is that the methods developed in this thesis research can be adopted on real application to interpret chart automatically. This interpretation result can be used as the second opinion to manual interpretation by the expert. The conclusion above is based on the user acceptance that reaches 80.1%, the percentage of keyword sufficiency reaching 66.7%, and the percentage of the dummy chart to occur in the real world reaching 93.6%.

5.2 Recommendations

For future works, the topic can be expanded to any other type of the chart such as bar charts, pie charts, etc (digitizer/semi automatic). In the future works, a multi dimensional or non-linear chart can be considered. The future researches also have to implement a fully automatic chart feature extraction for any chart (did not need to specify the type of the chart first). In the future, it is necessary to consider using a non-generated chart or manually drawing chart or using the scanned chart. According to the expert suggestion, the next research in Indonesian Language is recommended to use the words which has fuzzy meaning such as “kurang lebih”(more or less), “kira-kira” (about), or “sekitar” (about). Avoid using words “selalu” (always), “pasti” (surely), “tepatnya” (exactly), etc.

REFERENCES

[1] CCP14. (2010, July) Digitizing programs for Converting Hard Copy Graphs and Plots back to Data. [Online]. http://web.archive.org/web/20100728095151/http:/www.ccp14.ac.uk/solution/hardcopy2data.htm

[2] AIM, "Optical Character Recognition (OCR)," *The Association for automatic identification and data capture technologies*, 2000.

[3] the association for automatic Identification and data capture technologies, "Optical Character Recognition (OCR)," *the association for automatic Identification and data capture technologies*, no. 634 Alpha Drive Pittsburgh, PA 15238. Tel: +1.412.963.8588. Fax: +1.412.963.8753. Email: aidc@aimglobal.org. Web: www.aimglobal.org.

[4] Gary Huang, Carl Doersch, Erik Learned-Miller Andrew Kae, "Improving State-of-the-Art OCR through High-Precision Document-Specific Modeling".

[5] sourceforge. (2001, June) Plot Digitizer. [Online]. http://plotdigitizer.sourceforge.net/index.html

[6] Somayajulu Sripada, Jim Hunter, Jin Yu, Ian Davy Ehud Reiter, "Choosing Words in Computer-Generated Weather Forecasts," *elsevier, Artificial Intelligence*, vol. 167, pp. 137-169, 2005.

[7] Ehud Reiter, Ian Davy, Kristian Nilssen Somayajulu Sripada, "Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation," in *Proceedings of PAIS-2004, 2004.*, 2004, pp. 760-764.

[8] Somayajulu G Sripada and Feng Gao, "Summarizing Dive Computer Data: A Case Study in Integrating Textual and Graphical Presentations of Numerical Data," *CTIT Proceedings of the Workshop on Multimodal Output Generation*, pp. 149-157, 2007.

[9] Ehud Reiter, Jim Hunter, and Somayajulu G. Sripada Jin Yu, "A New Architecture for Summarizing Time Series Data," *In: Proceedings of INLG-04 Poster*, pp. 47-50, 2004.

[10] EHUD REITER, JIM HUNTER AND CHRIS MELLISH JIN YU, "Choosing the content of textual summaries of large time-series data sets," *Department of Computing Science, University of Aberdeen*, no. Natural Language Engineering, 2005.

[11] François Mairesse, "Natural Language Generation, ART on Dialogue Models and Dialogue Systems," *presentation on University of Sheffield*, 2006.

[12] Brian Roark and Eugene Charniak, "Noun-phrase co-occurence statistics for semiautomatic semantic lexicon construction," *In COLING-ACL*, p. pages 1110

[13] fdwiddows,beateg@csli.stanford.edu. Dominic Widdows and Beate Dorow, "A Graph Model for Unsupervised Lexical Acquisition," *Center for the Study of Language and Information. 210 Panama Street. Stanford University. Stanford CA 94305-4115.*

[14] Robert Dale, Mark Dras, Cecile Paris Stephen Wan, "Seed and Grow: Augmenting Statistically Generated Summary Sentences using Schematic Word Patterns," *Association for Computational Linguistics*, 2008.

[15] Kathleen R McKeown, "Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language," *Cambridge University Press*, 1985.

[16] Mirella Lapata, "Probabilistic text structuring: Experiments with sentence ordering," *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 545–552, 2003.

[17] Regina Barzilay and Lillian Lee, *Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization*, 113120th ed., Daniel Marcu Susan Dumais and Salim Roukos, Ed. Boston, Massachusetts, USA, May 2 - May 7.: HLT-NAACL, 2004.

[18] Fernando Agulló-Rueda. (2003–2005) http://www.icmm.csic.es/Fagullo/w3comput060.html. [Online]. http://www.icmm.csic.es/Fagullo/w3comput060.html

[19] Yuuki KABURAGI, Michiko SEO,Sumio TOKITA and Masao KANEKO Hidenobu SHIROISHI, "GetValue for Windows – Graph Digitizer Equipped with Electrochemical Analyzer," *J. Chem.*, vol. Vol. 8, no. No. 1, p. 37–40 (2002), January 2002.

[20] Andrew Kae, Carl Doersch and Erik Learned-Miller Gary B. Huang, "Bounding the Probability of Error for High Precision Optical Character Recognition.," *Journal of Machine Learning Research*, vol. 13, pp. 363-387, 2012.

[21] Eduard Hovy,.: University of Southern California, Marina del Rey, California, USA, ch. 4.

[22] Markus Dickinson, "N-grams," *Dept. of Linguistics, Indiana*, 2010.

[23] Hasan, dkk. Alwi, *Tata Bahasa Baku Bahasa Indonesia (Edisi Ketiga)*.: Jakarta: Balai Pustaka., 2000.

[24] Joice, *Pengembangan lanjut pengurai struktur kalimat bahasa Indonesia yang menggunakan constraint-based formalism*, 048765th ed. Depok: Fasilkom UI, 2002.

[25] Member, IEEE, Jens-Rainer Ohm, Member, IEEE, Vinod V. Vasudevan, Member, IEEE, and Akio Yamada. B. S. Manjunath, "Color and Texture Descriptors," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, vol. VOL. 11, no. NO. 6, JUNE 2001.

[26] Chad Dettlaff. Root Mean Square Error. PPT presentation.

[27] Sadao Kurohashi, Jun-ichi Nakamura XinYu Deng, "Building domain-independent text generation system," 2006.

[28] Timothée Poisot, "The digitize Package: Extracting Numerical Data from Scatterplots," *The R Journal*, vol. Vol. 3/1, no. ISSN 2073-4859, June 2011.

[29] Robert Dale Ehud Reiter, "Building Applied Natural Language Generation Systems," *Cambridge University Press*, vol. 1, no. Natural Language Engineering, 1995.

[30] Yuan-Fang Wang, Chan-Do Lee Chade-Meng Tan, "The Use of Bigrams to Enhance Text Categorization".