Indexing the Indonesian Web: Language Identification and Miscellaneous Issues

Vinsensius Berlian Vega S N
Department of Computer Science
National University of Singapore
3 Science Drive2, Singapore117543
vinsensi@comp.nus.edu.sg

Stéphane Bressan
Department of Computer Science
National University of Singapore
3 Science Drive2, Singapore117543
steph@comp.nus.edu.sg

ABSTRACT
Information retrieval tools and search engines have mainly been leveraging research results and technologies developed for the English language. In this paper we report the issues and obstacles we met in the process of designing and developing a search engine for the Indonesian language, as well as our progress and results. The results include original contributions such as a grammar for stemming Indonesian words and a self-improving language identification algorithm.

Keywords
Indonesian Language, search engine, web-crawler, stemming language identification, supervised learning, unsupervised learning.

1. INTRODUCTION
The Indonesian language, often referred to as Bahasa Indonesia, is the official language of the republic of Indonesia. Although several hundreds regional languages and dialects are used in the Republic, the Indonesian language is spoken by an estimated 200 million people, not counting an additional 20 million Malay speakers who can understand it. For a nation composed of several thousands islands and for its diasporas of students and professionals, the Internet, email, discussion groups, and the Web are unprecedented vehicles for cultural exchanges and for the conservation and development of an incomparably rich cultural diversity and identity. Yet few effective search engines are available to the Indonesian speaker wishing to search the Indonesian Web. The handful of Internet portals providing such a service relies on off-the-shelf technology designed for the American-English language.

We engaged in the design and development of a search engine for the Indonesian Web, i.e. the Web of documents written in Indonesian, with two objectives in mind. On the one hand, naturally, we aim at deploying our search engine. On the other hand, we identify and solve issues pertaining to the design and development of non-English language search tools and to create results of interest for other information retrieval or computational linguistic projects.

1.1 Search Engine Architecture
The search engine is composed of three main units, namely: the web-crawler, the indexing and retrieval modules, and the user interface. The crawler gathers documents across the web. It filters the Indonesian documents based on the language identification algorithm we describe in section 3. The indexing module processes the fetched documents: documents are segmented into words. Stop-words are removed. Remaining words are stemmed. The resulting terms are used to index the reference of the document (URL). Queries are processed similarly.

The retrieval module retrieves a ranked list of possibly relevant documents by comparing the queries with the documents. In the current implementation we use the SMART Information Retrieval system[1]. Although we ultimately aim at implementing our own retrieval engine, it is important to benchmark the other components using a reference test-bed such as SMART. The Information retrieval model used is the Vector Space Model [8].

1.2 Segmentation
The Indonesian language has officially adopted Roman alphabet. The large majority of Indonesian documents on the Web use the ASCII character set. The Indonesian language contains almost no diacritics except for some rare words assimilated from foreign languages. Dash "-", the numeral two "2", and the square symbol "^{2}" require a special handling. Indeed, plurals in Indonesian are expressed by repeating the noun (e.g. "buku-buku" = books), where the repeated noun can be adjoined by a dash. However, it is also a common practice to put the number 2 or the square symbol behind the word to denote repetition in writing (e.g. "buku2" or "buku^2"). The repeated forms have also evolved to indicate repetitive action (e.g. "jalan-jalan" = walking around) or other miscellaneous meaning (e.g. "mata-mata"= spy). To further complicate the matter, it is common to affix a repeated word or to repeat affixed words. Thus, certain mechanism needs to be developed to cater to this language feature. This feature is also the object corresponding stemming rules.

1.3 Stemming
The Indonesian language is a morphologically rich language. There are around 35 standard affixes (prefixes, suffixes, circumfixes, and some infixes inherited from Javanese) listed in [7]. Affixes can virtually be attached to any word and they can be iteratively combined. The wide use of affixes seems to have created a trend among Indonesian speakers to invent new affixes and affixation rules [2]. We refer to this larger set of affixes, which includes the standard set, as extended.

There are few implementations of stemming algorithms for the Indonesian language. Only one of which, from the University of Indonesia [6], was available for comparison. To our knowledge all existing algorithms use a dictionary and implement the standard set of affixes only.

By defining two sets of grammar rules corresponding to the derivation and inflection laws of Indonesian we constructed two algorithms for the stemming of the standard and extended sets of affixes, respectively. Our algorithms are based on the morphological rules only without dictionary.

[1] The SMART system (version 11.0) was developed at Cornell University and is available from ftp.cs.cornell.edu/pub/smart.

The performance of our stemming algorithm is comparable to that of the University of Indonesia's algorithm, when evaluated using the approach proposed in [4], i.e. from a computational linguistic point of view. In information retrieval, stemming is used to abstract from the morphological idiosyncrasies and hopefully results in an improvement of the retrieval performance. Our experiment with SMART showed again a performance of our algorithms comparable to the one of the dictionary-based algorithm. We noticed, however, that the performance increase over retrieval without stemming is only significant for queries involving common nouns and verbs (concepts) rather than proper nouns. As opposed to other morphologically rich languages such as Slovene [5], for which stemming brings a significant improvement of the retrieval performance, affixes in Indonesian are used to form derivations (conceptual variations) rather than inflections (grammatical variations). This refines the conclusion of [5] that the effectiveness of stemming is commensurate to the degree of morphological complexity in that we showed that it also depends on the role of the morphological rules.

2. IDENTIFYING INDONESIAN DOCUMENTS

The Indonesian Web is not a disconnected component of the World Wide Web. Web pages in Indonesian link to documents in English, Dutch, Arabic, or any other language. As we only wish to index Indonesian web pages, a language identification system that can tell whether a given document is written in Indonesian or not is needed. According to [3], language identification for text is a closed problem. Methods available yield near perfect performance. Among these methods, the most widely accepted are based on n-gram frequency [1]. However, all these methods differentiate documents from a set of known languages. This setting is unrealistic in the context of the web as one can neither know in advance nor predict the languages to be discriminated.

2.1 Learning from Positive Examples Only

We need to devise a language identification algorithm that can learn to distinguish between Indonesian and non-Indonesian documents from a reference set of Indonesian documents only. To put it in the Machine Learning context, we need an algorithm that learns from positive examples only. The algorithm we devised is based on the frequency of tri-grams in Indonesian words. The system first learns from a list of Indonesian words. Then, a weighted sum evaluates the similitude between the frequencies of the tri-grams in the reference set with those in candidate new document. A notion of penalty is also introduced to weight down tri-grams that have never been seen before. In our experiments we introduced three kinds of penalties. The algorithm achieves a performance of 94% recall and 88%precision.

2.2 Continuous Self-improvement from Self-labeled Examples

Collecting representative examples and deciding whether they are written in Indonesian or not for the initial training are time consuming tasks. Since the performance of the system was already very good we decided to try the bold idea of letting it be further iteratively trained by its own decisions.

We set an experiment with an initial training set of 9 Indonesian documents and a moderate penalty (to put the algorithm in mediocre initial and learning conditions). We measure the performance variation after each of 10 iterations. At each iteration, a set of 24 documents (12 Indonesian and 12 English) is presented to the algorithm, which would label and learn from the set. After each iteration, the performance of the algorithm is measured against a reference set (17 Indonesian, 4 English, 1 Malay, 1 Tagalog, and 1 German documents.) The performance of the algorithm (in terms of recall and precision) is shown in figure 1. The performance increase confirms the possibility of continuous self-improvement.

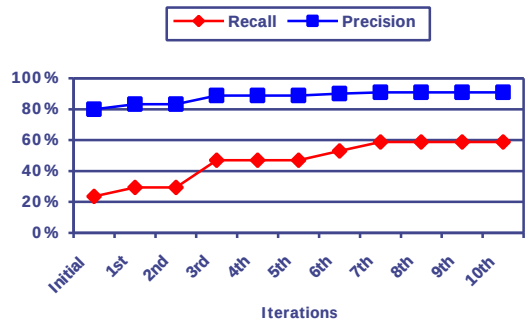

Figure 1. Recall and Precision for 10 iterations.

3. CONCLUSION

The World Wide Web is simultaneously an opportunity to foster a synergetic diversity of cultures and a risk to create a tame and dull global culture. We hope that our work and results on information retrieval for the Indonesian language can positively and constructively contribute to avoid a cultural and linguistic hegemony on the Web.

4. REFERENCES

[1] Cavnar, William B., Trenkle, M. *N-gram based text categorization.* Proceedings of the third Annual Symposium on Document Analysis and Information Retrieval, 11-13 April 1994, pp161-169.

[2] Kridalaksana, Harimurti., *Pembentukan Kata Dalam Bahasa Indonesia.* P.T. Gramedia, Jakarta 1989.

[3] Lazzari, G., et all. *Speaker-language identification and speech translation.* Part of Multilingual Information Management: Current Levels and Future Abilities, delivered to US Defense ARPA, April 1999.

[4] Paice, Chris D., *An evaluation method for stemming algorithms.* Proceedings of the 17th annual International ACM-SIGIR conference on Research and Development in Information Retrieval 1994, pp 42-50.

[5] Popovic, Mirko., and Willett, Peter., *The effectiveness of stemming for natural-language access to Slovene textual data.* Journal of the American Society for Information Science, Vo. 43, June 1992, pp 384-390.

[6] Siregar, Neil Edwin F., *Pencari Kata Berimbuhan pada Kamus Besar Bahasa Indonesia dengan Menggunakan Algoritma Stemming.* Final Year Thesis, University of Indonesia 1995.

[7] Tim Penyusun Kamus, *Kamus Besar Bahasa Indonesia.2ed.* Balai Pustaka, 1999.

[8] Yates, R. B. and Neto, B R., *Modern Information Retrieval.* ACM Press New York, 1999.