Word-based Semantic Computation for Multilingual Information Retrieval in the Context of Emotion Expressions

Student Name: Totok Suhardijanto (Doctor Student 2nd year)
Supervisor: Prof. Kiyoki Yasushi
Student ID: 80849341

Abstract

In this research, we develop a system to provide user not only with regular information, but also additional information such as feeling, mood and emotion. Recently, with the transition of the computer platforms from a device with limited function into an omnipotent media, other functions such as extracting emotional information, recognizing feelings, or revealing ideological orientation become more and more worth to be invented. Along with this transition, this research proposed a system to offer user a function to extract and recognize emotional information in multicultural environments.

In this research, we made use of vector space model to measure and compute the similarity and correspondence across languages and support the system with one space for one language and specific space for specific function. In our experiment, we implemented the system to retrieve emotion expressions in several languages simultaneously. We used collections of emotion expressions in Japanese, English, Filipino and Indonesian we defined based on the semantic atlas of emotion conceptual [9].

Background

The intense communication between Internet users with various cultures in the world currently makes the World Wide Web becomes more multicultural. In the previous days, English had dominated WWW, but in current situations, there are several languages in the Internet becoming major languages of the Internet such as Arabic, Chinese, Malay/Indonesian, and Japanese. An efficient and robust multilingual information retrieval and data mining are now broadly developed in order to provide a better way and approach to support cross-cultural and multicultural activities in the Internet. On other hand, with the advantage of modern computer that allow user to exchange not only textual information, but also multimedia information, modern computer also face problems in dealing with not only primary or textual information, but also secondary or additional information such as affection, tone, orientation and attitude. In response to this orientation in the area of Information Retrieval, we proposed a new system of Multilingual Information Retrieval with function of emotion processing.

Multilingual Information Retrieval (MLIR) and Cross-Languages Information Retrieval (CLIR) are very challenging and attracting in the area of information retrieval. Many MLIR and CLIR systems have been proposed and widely applied for both commercial purposes and research systems. The central issue in MLIR/CLIR is that the input is in one language (L1) and the output will be provided in another language (L2) [5]. In the case of MLIR, the system usually translates query, or target-document/media, or both query and document/media into a pivot language to unify the language of query and document/media. The other critical point in MLIR/CLIR is whether the query only consists of one keyword, phrase, or sentence [6]. MLIR with one keyword-query usually has problem in obtaining retrieved results with higher precision rate. On the other hand, MLIR system that accommodates phrase or sentence-query enables to raise the precision rate, but the system itself is quite complicated and, to some extent, it makes the computing time takes longer. We implemented semantic computation using vector space model in combination with the Mathematical Model of Meaning (MMM) to make the system simpler and to accelerate computation time. We also distributed specific application and function into specific space in order to give a more precisely result. In the conventional system, the query is in L1 and the result will be provided in another language (L2). In this system, we introduced a novel system that provides results in multi-languages (L2 ... Ln).

Objective

In this research we developed an Emotion-based Multilingual Information Retrieval system with an automatic translation to retrieve document/other media in languages different from the language of query. This system realized a retrieval engine for multilingual information with query-translation. We based our method on the linguistic theory in the matter of that most of words in some language cannot be translated into one to one correspondence to those in other languages [7], [8]. This approach made our system to be more similar to cross-cultural communication and interaction in natural way because in the natural language a word in one language cannot totally be translated into a word in another language. Figure 1 illustrates the core engine in the proposed system.

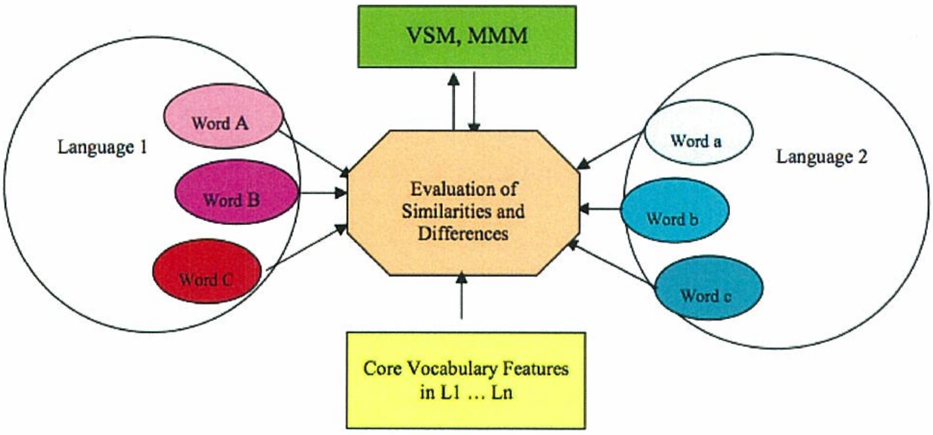
VSM, MMM
Language 1
Word A
Word B
Word C
Evaluation of Similarities and Differences
Word a
Word b
Word c
Language 2
Core Vocabulary Features in L1 ... Ln

Figure 1 Core Engine of Emotion-based Multilingual Information Retrieval (EMLIR) System)

In our system, a specific semantic space was created for each language. For semantic space creation, in this research we used 2114 core vocabulary of English based on the Longman Dictionary of Contemporary English (2006) [1]. For other languages, we used different number of core vocabulary based on recent survey and study in the languages [2], [3], [4], [10], [11], [12], [13]. Based on these core vocabularies, the system extracted features using dictionaries in each language automatically. And then, the core vocabularies and features are written into matrix to create semantic metadata space and being converted as vector. The retrieval system translates a query directly into target languages by applying semantic associative space search engine. Query in L1 is submitted into front-end controller, and then it will be translated by using semantic associative space and MMM.

The metadata of query in L1 is computed in order to measure the similarity with metadata repository of target multimedia database in L2. Target multimedia result in L2 are separately ranked and indexed in each language. For the text-document retrieval, the result lists are merged and presented in a multilingual list of text-documents. In this research we apply a multi-centralized merging approach to integrate the result lists of collections in different languages. The ranked list is then used to provide the more precisely output. For other media retrieval, the result is presented in each language. It means that the result is provided into several lists related to each language/culture. Figure 2 shows the system architecture of the proposed system and Figure 3 shows the further application of our core engine.

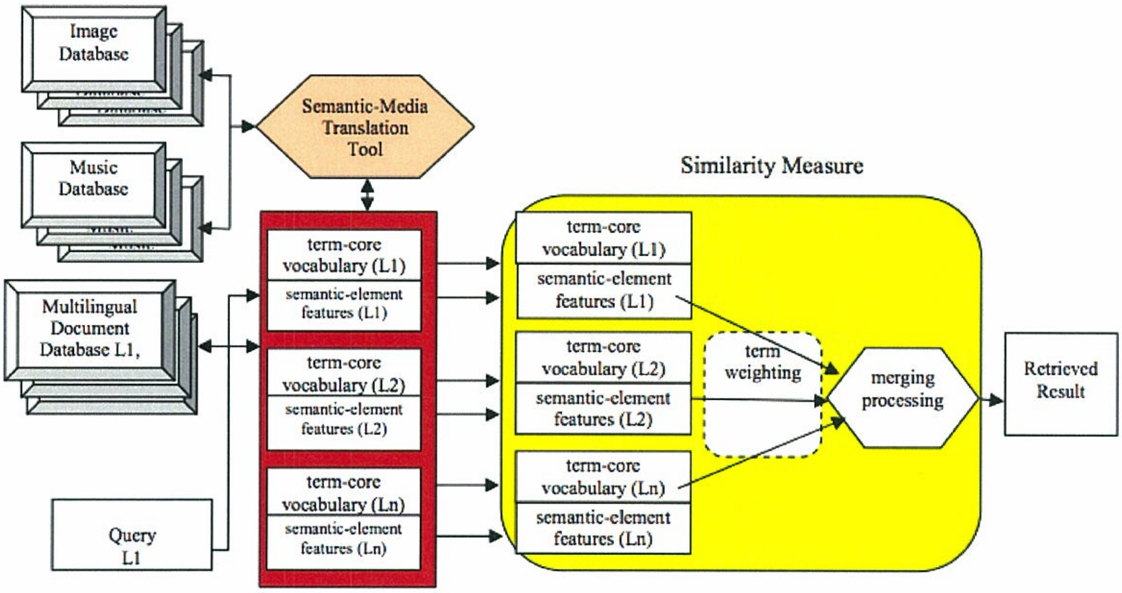
Image Database
Music Database
Multilingual Document Database L1,
Query L1
Semantic-Media Translation Tool
term-core vocabulary (L1)
semantic-element features (L1)
term-core vocabulary (L2)
semantic-element features (L2)
term-core vocabulary (Ln)
semantic-element features (Ln)
Similarity Measure
term weighting
merging processing
Retrieved Result

Figure 2 System Architecture of EMLIR System

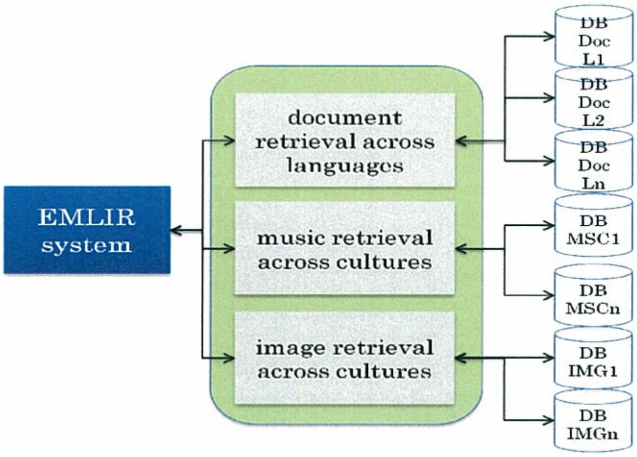
EMLIR system
document retrieval across languages
music retrieval across cultures
image retrieval across cultures
DB Doc L1
DB Doc L2
DB Doc Ln
DB MSC1
DB MSCn
DB IMG1
DB IMGn

Figure 3 Application Possibility for EMLIR System

Result

In this research, we developed a multilingual information retrieval system that dealing with multimedia database feature in 4 major languages of the world in which three of them are Asian top languages according to Internet World Stats (http://www.internetworldstats.com/stats3.htm#asia). This system included three core components, including semantic space, semantic-media translation, and automatic weighting. In the semantic space, a query is translated in the following method: (i) extracting its features, and (ii) computing them to retrieve keywords in L2, L3, and L4. For text-document retrieval, the retrieved keywords are used to call related documents in L2, L3, and L4 document-database. As for multimedia retrieval, retrieved keywords in L2, L3 and L4 are converted into image or music feature by using semantic-media translation to retrieve targeted images or music from multimedia database. An automatic weighting then compute and analyze retrieved files by comparing them to the query features. The output of this system will be provided into lists of documents, images or music that is related to the query. In this system, the output of text-document retrieval is a list of documents written in multi-languages; the output of image and music retrieval is a list of image or music that is defined by features of other cultures or languages.

5. References

[1] *Longman Dictionary of Contemporary English*, 2006.

[2] *Longman: A Comprehensive of Indonesian-English*, 2004. Ohio: Ohio University Press.

[3] *Kenkyusha's New Japanese-English Dictionary*, 2003. Tokyo: French & European Pubns.

[4] *Webster's Tagalog-English Thesaurus Dictionary*. 2008. Massachusetts: BookSurge Publishing.

[5] Lin, Wen-Chen and Chen, Hsin-Hsi. 2002. Merging Mechanism in Multilingual Information Retrieval. *Cross-language Evaluation Forum* (CLEF), Rome, Italy, September 19-20, 2002.

[6] Abdelali, Ahmed, Cowie, James , Farwell, David and Ogden, William. 2002. UCLIR: a Multilingual Information Retrieval Tool. *Multilingual Information Access and Natural Language Processing* (IBERAMIA 2002), Sevilla, Spain, September 12, 2002.

[7] Yamamoto K., and Y Matsumoto. 2000. Acquisition of Phrase-Level Bilingual Correspondence using Dependency Structure. In Proceedings of COLI□G 2000, Saarbrueken, Germany, pp. 933-939.

[8] Vinay, Jean-Paul and Darbelnet, Jean. 2000. *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam: John Benjamin Publishing.

[9] Averill, J. R. (1975). A semantic atlas of emotional concepts. *JSAS: Catalog of Selected Documents in Psychology*, 5, 330. (Ms. No. 421).

[10] Quinn, George. 2001. *The Learner's Dictionary of Today's Indonesian*. Sydney :Allen & Unwin.

[11] Akiyama, Carol. 1991. *Japanese Vocabulary*. New York: Barron's Educational Series.

[12] Shoji, Kakuko. 2001. *Japanese Core Words and Phrases: Things You Can't Find in a Dictionary* (*Power Japanese Series*). Tokyo: Kodansha.

[13] Zorc, R. David Paul. 1979. *Core etymological dictionary of Filipino*. Darwin: Darwin Community College.