2008－MUS－78 (8)
2008／12／19

音楽情報検索向け類似テキスト検索システムの試作

徐昕 内藤正樹 加藤恒夫 河井恒
株式会社 KDDI 研究所

本稿では、音響的な距離を考慮することによって、誤りを含むクエリであっても検索可能にした歌詞による音楽情報検索システムを試作したので報告する。従来全文検索システムでは、聞き間違い等により、検索クエリとしての歌詞が登録されている歌詞と完全一致しない場合には検索成功率が大きく低下する問題があった。提案したシステムでは、文字列をマッチングする際に単語間の音響的な距離を考慮するため、歌詞が完全一致しない場合でも音響的に近い歌詞であれば検索に成功する。検証実験により、提案システムが聞き間違えた歌詞による検索時の精度向上に有効であることが判明した。

An Introduction of a Fuzzy Text Retrieval System For Music Information Retrieval

Xin XU, Masaki Naito, Tsuneo Kato, Hisashi Kawai
KDDI R&D Laboratories

This paper proposes a new fuzzy text retrieval system for music information retrieval by queries of lyric phases. The differences between queries and retrieval objects (the lyric words) critically deteriorate search accuracy of Web search engines and full-text retrieval systems. The proposed retrieval system improved search accuracy by taking account of the acoustic similarity during string matching process, especially in the case that users mishear the words in the input queries.

1. Introduction

Recent popularity of online music distribution services, such as iTunes Music Store, LISMO, have enabled common users to download and enjoy more digital music resource. The music distribution business income of 2007 in Japan reaches about 755 billion yen [1].

An effective music information retrieval (MIR) system plays an essential role for realizing satisfactory music distribution services, and improving usability of music applications. Many music services are providing MIR systems which accept queries by text, humming, and singing voice. Among them, MIR by text still stays in the main position by a faster response and higher search accuracy.

In this paper, we focus on music information retrieval by queries of lyric words. This task is realized by a fuzzy text search system with lyric database. In a lot of cases, users remember the lyric words when they are impressed by hearing a part of a song without a lyric sheet. With an ambiguous human memory, it is not practical that user can input the correct lyric words every time. Also, typos and different transcriptions are not rare in lyric database, which also deteriorates search accuracy. The authors' preliminary investigations on real world queries suggested that acoustic confusion causes mistakes in queries. In order to retrieve lyric texts or other music information by these queries, a

robust fuzzy text retrieval system is necessary. So far, some algorithms have been proposed to deal with fuzzy text retrieval, such as LSI (Latent Semantic Indexing) [3] and string metric of edit distance [4]. However, they don't help in the cases that the query words are mistaken by acoustic confusion. To solve this problem, we propose a fuzzy text retrieval system, which takes account of acoustic similarity during string matching process. An indexing method is also implemented to make the on-line search processing efficient.

The remainder of this paper is organized as follows: we describe how the mistaken queries are collected and analyzed in Section 2. The proposed system is introduced in Section 3. In Section 4, the proposed system is evaluated with other search systems. The conclusion is given in the end.

2. Collection and Analysis of Lyric Query Mistakes

To analyze the actual query texts of lyric words for MIR, we collected real word queries that are asked by users in a question & answer community site "Oshiete goo! (教えて goo!)". According to the answers provided by other users, 283 correct lyric texts can be known. The collected queries and the correct lyric texts are compared to analyze how users mistake the lyric words.

The analysis results are shown in Figure 1 and Table 1. Figure 1 presents the distribution of mistaken queries of different types and queries without errors within the collected queries. Almost half of the queries are mistaken in the content word (noun, verb, adj etc...). The standard full-text retrieval system based on a complete match rule fails to search this kind of queries. On the other hand, the queries of function-error-word type are usually handled by the set of stop words in retrieval systems. Among the content-word-error queries, the reason of mistake can be categorized in 3 types, which are "acoustic confusion", "meaning confusion", and "others". The percentages and examples are listed in Table 1. As the mistaken parts are marked in bold, the meaning confusion is defined that a word is mistaken by another word which has a similar meaning or is in the same semantic category. The acoustic confusion is regarded as the word is mistaken by another word which has a similar pronunciation. The type of "others" contains word insertion, word deletion, and other errors in queries. Since the occurrences of "others" are too difficult to be summarized, the discussion of it is not included in this paper.

Some algorithms of fuzzy text retrieval are considered to be solutions against searching failures caused by first two mistaken queries.

LSI is an approach to take advantage of implicit semantic structure to determine what a page is about outside of specifically matching search query text. It means LSI considers documents that have many words in common to be semantically close, and ones with few words in common to be semantically distant. LSI has been implemented into many main web search engines, such as Google. As an example, for "Apple", by LSI technology, it expects to find the terms Apple, Windows and Mac on the page. In the same way, the mistaken queries of meaning confusion in Table 1 are input into Google search site, and then the pages related to the target lyric texts come up.

Also some searching methods use string metric, such as edit distance, to find similar strings in a string set.

However, the acoustic confusion problem, as the example is shown in Table 1, " inori(祈り, prey)" and

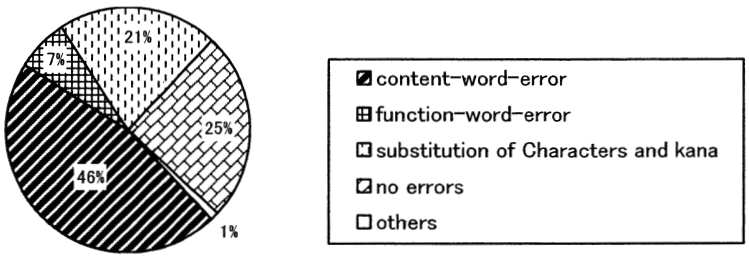

Figure 1. The distribution of mistaken queries in different types and queries without errors within the collected queries

Table 1. The distribution of mistaken types within Content-word-error cases

Types of errors	Percentage	Correct lyric (examples)	Mistaken queries (examples)
meaning confusion	41%	君(you)には何でも話せるよと	あなた(you)には何でも話せるよと
		月(moon)に願いを	星(star)に願いを
acoustic confusion	38%	遠い海にひとり(one person)	遠い海に祈り(prey)
		無数(countless)に延びる放射状の光	まっすぐ(straight)に伸びる放射状の光
others	21%	星から来た(come from)子の見る夢は	星の子チョビン(chobin)の見る夢は

Table2. A sample of index item

Lyric text NO.:10				
Preset keyword list	Acoustic candidate	Acoustic distance	Character edit distance	TFIDF
若者(youth)	赤帽(red hat)	0.4	1	0
...	...	...	...	...
熱情(passion)	熱情(passion)	0	0	0.4

"hitori(ひ と り , alone)" are similar in pronunciation, while are distant in semantic and character string. It cannot be solved by two algorithms introduced above. To find a solution for acoustic confusion, we propose a system which takes acoustic distance into the fuzzy text search processing.

3. System Structure

The proposed retrieval system comprises indexing part and searching part.

Indexing part:

First, each lyric text *Ti* of the lyric database is segmented by a word segmenter to generate a list of words. Here we use Mecab [5] as the segmenter. Accompanying the segment process, a word filter is used to get rid of stop words that belong to unimportant parts of speech (article, conjunction, prep). After the process, a word sequence *Si* is obtained from each lyric text *Ti*.

Second, a preset keyword list is created, which is assumed to contain all possible words in input queries. The words are collected from lyric database. Also in order to cope with the mistaken words which

may not belong to the lyric domain, words from other domains are added.

Third, for every word *An* in the preset keyword list, a most acoustically similar word candidate *A'i,n,* which has the minimum acoustic distance with *An*, is selected from each word sequence *Si*. The acoustic distance *ADi,n* between *An* and *A'i,n* is recorded in the index file, as shown in Table 2.

The acoustic distance is calculated as follows:

The entire set of words is transcribed into phone sequences. Then, DP (Dynamic Programming) matching is used to calculate the acoustic distance between two phone sequences. The insertion and deletion penalty are set to 1. The substitution penalty is set to Mahalanobis distance [2] between two phones' acoustic models. Mahalanobis distance *MDi,j* between phone *i* and phone *j* is defined in Eq.(1) [2].

$$MDi,j = \sqrt{\frac{K\sum_{k=1}^{K}(\mu_{ik}-\mu_{jk})^2}{\sum_{k=1}^{K}\sigma_{ik}^2+\sum_{k=1}^{K}\sigma_{jk}^2}} \qquad (1)$$

where, K is the number of acoustic feature vector dimensions of the acoustic models. Here, acoustic features are MFCC (Mel Frequency Cepstral Coefficient), their 1st and 2nd derivatives, and 1st and 2nd derivatives of the power. *K* is 38. μik and σik are *kth* order mean and variance. Single mixture monophone models with 5 states are used in Mahalanobis distance calculation.

In addition to acoustic distance, other parameters for ranking, such as, character edit distance *CD* and *TFIDF* (term frequency–inverse document frequency) are calculated. Character edit distance *CDi,n* is the minimum of edit distance values between *An* and each word in *Si*. The penalties of insertion, deletion, substitution are all set to 1 in the calculation of edit distance. *TFIDFi,n* of the candidate word *An* is calculated by Eq. (2).

$$TFIDF_{i,n} = \begin{cases} 0, & if \quad AD_{i,n} \neq 0 \\ tf_{i,n} \times \log(\frac{N}{df_n}) & if \quad AD_{i,n} = 0 \end{cases} \qquad (2)$$

$tf_{i,n}$ is the number of occurrences of *An* in *Ti*, and df_n is the number of lyric texts containing *An*. *N* represents the total number of lyric texts.

For each lyric text *Ti*, an index term is created with preset keyword list, acoustic candidate *A'i,n,* the values of the acoustic distance *ADi,n*, character edit distance *CDi,n* and *TFIDFi,n*, as is shown in Tabe 2.

Searching part:

The flowchart of searching part is illustrated in Figure 2.

First, the input query is segmented and filtered into a word sequence *Q1, Q2, ..., Qm* in the same way with the indexing part.

Second, *Q1, Q2, ..., Qm* are located in the preset keyword list, e.g., *Q1=A3, Q2=A9, ..., Qm=An.*

Referring to the information recorded in the index file, the ranking score *Ri* of each lyric text *Ti* is calculated by Eq.(3).

$$R_i = \alpha \sum_{n=q_1,..q_m} AD_{i,n} + \beta \sum_{n=q_1,..q_m} CD_{i,n} + \gamma \sum_{n=q_1,..q_m} (1 - TFIDF_{i,n}) \qquad (3)$$

The values of *ADi,n, CDi,n, TFIDFi,n* are normalized from 0 to 1. The weights α, β, γ are set to 1, 0.5,1 respectively, based on our preliminary experiments.

Because the lyric text *Ti* with lower ranking score *Ri* is regarded as more suitable candidate, 1−*TFIDFi,n* is used instead of *TFIDFi,n*.

4. Experiments

To evaluate search accuracy, the proposed method is compared with some traditional methods. A

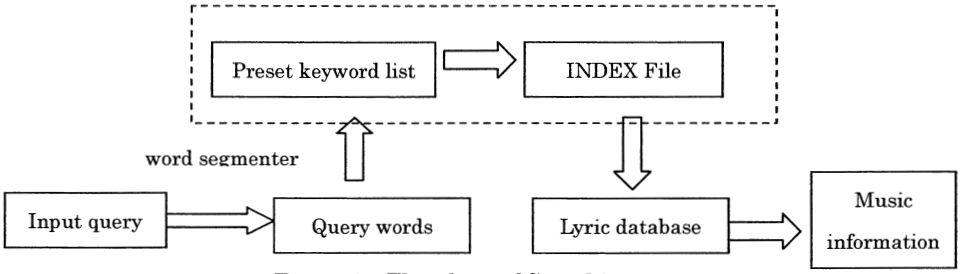

Figure 2 Flowchart of Searching part

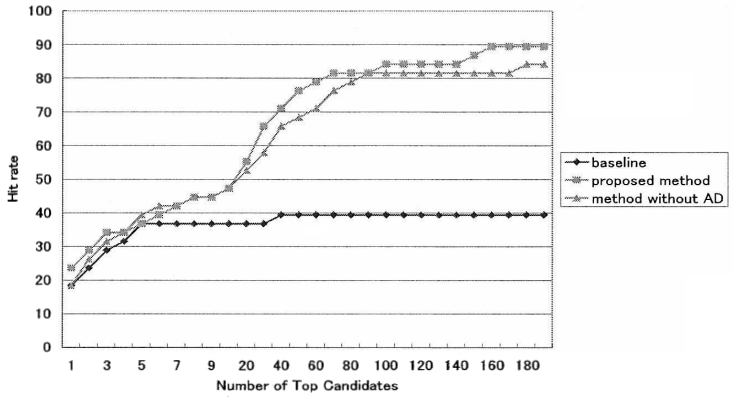

Figure 3 Comparison results of Hit rate-Number of Top Candidates by "mistaken queries"

Table 3 Hit rate within top 20 candidates

	Google Search	Baseline	Proposed method
mistaken queries	10.1%	36%	55%
correct queries	73%	100%	100%

database of 3743 lyric texts is collected. It contains both Japanese and English lyrics. The test set contains 2 groups of queries, "mistaken queries" and "correct queries". "mistaken queries" consist of 38 queries from the collected mistaken queries. "correct queries" represents correct lyric parts in the lyric texts that correspond to 38 mistaken queries.

The first experiment is carried out to evaluate how much the acoustic distance affects the rank of target lyric file text. It uses the "mistaken queries" test set. The result of 'Hit rate-Number of Top Candidates' is shown in Figure 3, in which the number of top candidates is ranged from 1 to 190.

"method without AD" means the method with the same algorithm as the proposed method only except the acoustic element. The ranking score Ri' of "method without AD" is calculated by Eq.(4). From Figure 3 we can see that, when the number of top candidates is from 1 to 10, two methods achieve almost the same hit rates. On the other hand, in the part where the number of top candidates is from 20 to 190, the proposed method improves the hit rate by 3% ~ 8%, comparing with "method without AD".

It means that during this part, the ranks of target lyric text are raised by the calculation of acoustic distance. By analyzing the mistaken queries whose hit rates are improved by the proposed method, it is found that all of them belong to the type of acoustic confusion.

$$R'_i = \sum_n CD_{i,n} + \sum_n (1 - TFIDF_{i,n}) \qquad (4)$$

In the second evaluation, 3 retrieval methods, a baseline, "Google Search", and the proposed method are compared. The baseline is a text retrieval system of inverted index algorithm with TFIDF ranking weight. It is a fundamental method of Web search engine and full-text retrieval system. "Google Search" method inputs queries into Google web search engine, then check whether the target music information (lyric text) is contained in the hit pages. Considering the search experience in real word, hit rates within top 20 candidates are used to evaluate the search accuracy. The experimental results are shown in Table 3. The proposed method outperformed the other two methods, which are drastically degraded by the mistaken queries.

5. Conclusion

This paper introduced a fuzzy text retrieval system that takes account the acoustic similarity during string matching process. According to the evaluation results, the proposed system keeps robust against the mistaken queries while the search accuracies of Google search and baseline are dramatically deteriorated. However, comparing with the method without using acoustic distance, the proposed method doesn't improve the search accuracy of mistaken queries much, when the number of top candidates is below 10. To further raise the ranks of target lyric texts corresponding to the mistaken queries, an optimized position parameter based on the positions of the words in queries and in lyric texts is necessary. It remains as an essential task in our future research.

Reference:

[1] http://www.riaj.or.jp/data/download/2007.html

[2] 中村 匡伸, 岩野 公司, 古井 貞熙 "マハラノビス距離を用いた日本語話し言葉音声の音響的特徴の分析" 日本音響学会 2005 年春季講演論文集, 2-1-4, pp.230-231 (2005)

[3] S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. "Indexing by Latent Semantic Analysis". Journal of the American Society for Information Science 41(6), pp.391-407 (1990)

[4] 花田 博幸, 工藤 峰一,"編集距離による最類似文字列の探索高速化に関する研究",電子情報通信学会技術研究報告, Vol.108, No.93,pp. 41-45 (2008)

[5] http://mecab.sourceforge.net