Journal of Digital Contents Society Vol. 14 No. 4 Dec. 2013(pp. 463-469)
http://dx.doi.org/10.9728/dcs.2013.14.4.463

중복글자 구분을 지원하는 집합 기반 POI 검색 알고리즘 구현

고은별*, 이종우**

요 약

집합 기반 POI 검색 기법은 부정확한 질의어 입력 시 검색결과의 재현율과 정확도가 현저히 떨어지는 기존 하드매칭 기법에 비해 우수한 성능을 보인다. 하지만 집합 개념을 바탕으로 했기 때문에 한 POI 레코드 내에 중복으로 포함된 동일 글자들을 구분하지 못하는 문제점이 있다. 본 논문에서는 이러한 문제를 해결하여 한 POI 내에 동일 글자가 여러 개 등장해도 동작하는 개선된 집합 기반 POI 검색 기법을 제시하고 구현하였다. 또한 개선된 집합 기반 POI 검색 기법의 검색결과와 기존 집합 기반 POI 검색 기법의 검색결과를 비교하는 실험을 통해 레코드 내에 중복으로 포함된 동일 글자가 있는 레코드에 대한 검색 성능이 향상되었음을 확인하였다.

키워드 : POI 데이터베이스, POI 검색 알고리즘, 집합-기반 알고리즘, 중복 글자

Implementation of A Set-based POI Search Algorithm Supporting Classifying Duplicate Characters

Eunbyul Ko*, Jongwoo Lee**

Abstract

The set-based POI search algorithm showed better performance than the existing hard matching search when inaccurate queries are entered. In the set-based POI search algorithm, however, there is a problem that can't classify duplicate characters within a record. This is due to it's 'set-based' search property. To solve this problem, we improve the existing set-based POI search algorithm. In this paper, we propose and implement an improved set-based POI search algorithm that is able to deal duplicate characters properly. From the experimental results, we can find that our technique for duplicate characters improves the performance of the existing set based POI search algorithm

Keywords : POI database, POI search algorithm, Set-based algorithm

1. 서론

스마트폰이 대중화되면서 위치 기반 서비스를 이용한 지도/네비게이션 어플리케이션이 활성화되고 있다.[1] 과거 지도/네비게이션에 대한 활발한 연구의 결과로[2] 현재는 언제 어디서나 가고자 하는 목적지를 입력하면 원하는 정보를 얻을 수 있다. 또한 추출된 정보를 사용자가 이해하기 쉽게 전달하기 위한 인터페이스 연구도 활발하다.[3,4,5] 하지만 사용자가 목적지의 명칭을 정확하게 입력하지 못하는 경우가 많아 목적지 검색의 재현율과 정확률이 떨어진다. 전통적인 하드매칭 기법에서는 이런 부정확한 질의어를 제대로 처리하지 못해 검색 서비스의 성능 저하 문제를 야기한다. 이 문제를 해결하기 위해 집합 기반 POI 검색

※ 교신저자(Corresponding Author): Jongwoo Lee
접수일:2013년 11월 18일, 수정일:2013년 12월 20일
완료일:2013년 12월 25일
* 숙명여자대학교 멀티미디어과학과
** 숙명여자대학교 멀티미디어과학과 교수
Tel: +82-2-710-9952 , Fax: +82-2-710-9704
email: bigrain@sm.ac.kr
▣ 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2013R1A1A2013155)

기법[6]에서는 의미있는 실험결과를 보인 바 있다.

하지만 집합 기반의 POI 검색 기법은 집합 개념을 바탕으로 하기 때문에 한 레코드 내에 동일 글자가 2개 이상 있는 경우를 처리하지 못한다. 같은 글자가 레코드에 여러 개 포함되어 있어도 집합 개념 검색이므로 단 한 개만 포함되어 있는 것으로 처리하는 것이다. 그래서 이런 중복 글자가 있는 명칭을 질의어로 입력하면 검색 정확도가 현저하게 떨어진다. '동성자동차'처럼 같은 글자를 포함하고 있는 질의어가 입력될 경우 중복된 '동'을 구별하지 못하고 하나로 보기 때문에 '한성자동차', '성원자동차'같은 결과가 검색 일치로 나오게 된다. 문자열 유사도 측정을 통해 이 같은 문제를 일부 보완했지만 근본적인 해결책은 되지 못한다.[7,8]

이런 문제를 해결하기 위해 본 논문에서는 기존 집합 기반 POI 검색 기법을 보완한 중복 글자 검색을 지원하는 집합 기반 POI 검색 기법을 제안한다. 동일 글자가 한 레코드 내에 여러 번 등장할 경우 최대 등장 횟수만큼 글자 아이디를 더 생성한다. 동일 글자이지만 글자별로 부여하는 아이디를 달리 하는 것이다. 레코드 내 중복을 반영한 글자 아이디는 역 인덱스 생성에 반영된다. 그래서 기존 집합 기반 POI 검색 기법에서는 인식하지 못했던 중복 글자들을 정확하게 인식하여 검색 성능을 향상시킨다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 집합 기반 POI 검색 기법에 대해 설명하고, 3장에서는 기존 집합 기반 POI 검색 기법의 한계에 대해서 설명한다. 4장에서는 본 논문에서 제안하는 중복 글자 지원 집합 기반 POI 검색 기법을 제시하고, 5장에서는 본 논문에서 제시된 알고리즘의 실효성을 보인 뒤, 6장에서 결론을 맺는다.

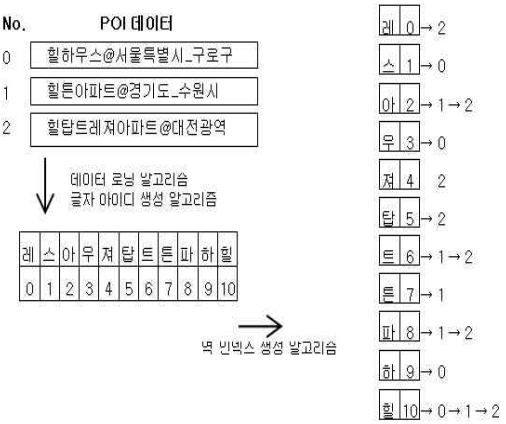
2. 기존 집합 기반 POI 검색 기법 소개

본 장에서는 기존 집합 기반 POI 검색 기법에 대해서 간략히 설명한다. 집합 기반 POI 검색 기법[6]이란 집합 개념을 적용하여 POI 데이터를 검색하는 기법으로, 전역적 쿼리 확장[9,10], 지역적 쿼리 확장[9,10], 적합성 피드백[11]과 같은 계산이 복잡하고 방대한 양의 데이터를 필요로 하는 기법을 사용하기 힘든 차량 내비게이션과 같은 독립형 시스템을 위한 기법이다. 부정확한 POI 질의어 입력으로 인한 POI 검색 서비스 성능 저하 문제를 시스템 자체 내에서 해결하기 위해 제시한 새로운 알고리즘이다. 검색 과정을 간략히 소개하면 다음과 같다.

기존 집합 기반 POI 검색 기법은 POI 디비 로딩 알고리즘, 글자 아이디 생성 알고리즘, 역 인덱스 생성 알고리즘, 텍스트 검색 알고리즘으로 구성되며, 텍스트 검색 알고리즘은 전처리 과정, 블록 내 연산, 블록 간 연산으로 구성된다.

먼저, POI 디비 로딩 알고리즘에서는 POI 데이터베이스를 로딩하며, 글자 아이디 생성 알고리즘에서는 레드-블랙 트리를 이용하여 로딩된 데이터베이스의 모든 글자를 중복없이 파악하고 이를 통해 각 글자의 아이디를 생성한다. 이 아이디는 역 인덱스 생성 알고리즘에서 사용된다. 역 인덱스에는 아이디가 부여된 각 글자가 포함되어 있는 POI 데이터베이스 내 레코드 번호를 저장한다. (그림 1)은 세 개의 레코드에 대해 글자 아이디와 역 인덱스를 생성하는 과정 예를 표현한 것이다.

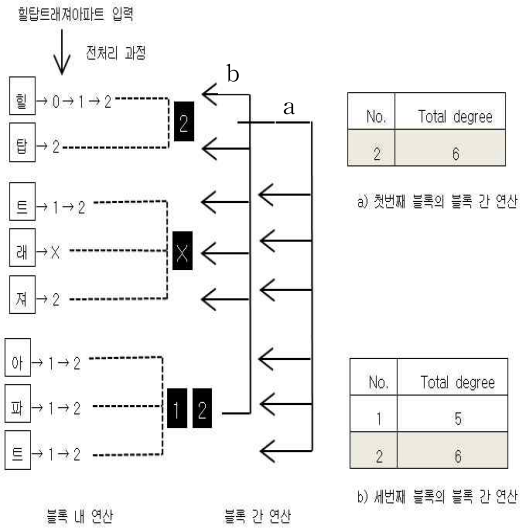
(그림 1) 역 인덱스 생성 과정



(Figure 1) Steps for generating inverted indexes

이렇게 생성된 역 인덱스는 텍스트 검색 알고리즘을 통해 검색에 이용된다. 먼저, 길이 n인 질의어를 k개의 블록으로 균등 분할하여 블록 내 연산을 수행한다. 역인덱스를 활용하여 각 블록의 글자를 모두 포함하고 있는 레코드를 파악한다. 그 후 블록 간 연산을 통해 다른 블록의 정보를 활용하여 질의어와 가장 근접한 POI 데이터베이스 내 레코드를 출력한다. 이때 '차수'라는 개념을 사용하는데, 차수란 주어진 레코드에 포함된 질의어 내 글자의 총 개수이다. 이러한 텍스트 검색 알고리즘의 동작 과정은 (그림 2)와 같다.

(그림 2) 텍스트 검색 과정



(Figure 2) Steps for searching a text

3. 기존 집합 기반 POI 검색 기법의 한계

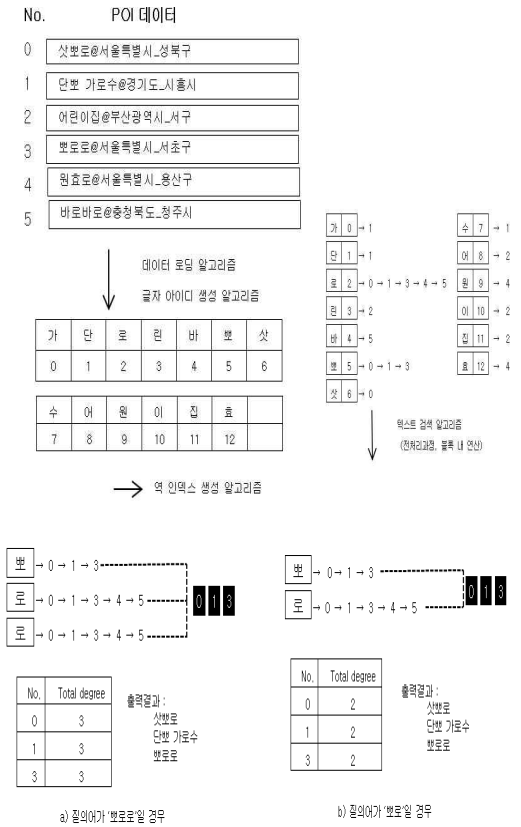
기존 집합 기반 POI 검색 기법은 기존의 하드매칭 기법을 사용하는 시스템에 비해 검색 성능을 현저히 높였다는 장점이 있었고, 또한 집합 개념 검색을 사용했으므로 질의어 문자들과 연속 일치하지 않는 결과들도 출력해 줌으로써 검색이 전혀 안 되던 오질의어에 대한 검색율도 획기적으로 향상시켰다. 하지만 집합 개념을 바탕으로 했기 때문에 한 레코드 내에 중복으로 포함된 동일 글자들을 구분하지 못하는 문제점이 있다. 이 장에서는 기존 집합 기반 POI 검색 기법의 이 같은 한계에 대해 설명한다.

기존 검색 기법에서는 레드-블랙 트리를 이용하여 중복을 제거한 데이터베이스의 모든 글자에 아이디를 부여한다. 이 아이디를 이용하여 역 인덱스 생성 시 해당 글자가 포함된 레코드 번호들만을 저장해둔다. 그렇기 때문에 해당 글자가 한 레코드에 몇 번 등장하는지는 전혀 고려하지 않는다. 이렇게 생성된 역 인덱스는 중복글자가 있는 질의어를 입력받았을 때 중복된 각 글자들을 구분하여 등장 줄 번호들을 알려주지 못한다는 한계가 있다.

앞에서 예로 든 '힐탑트래져아파트'도 첫번째 '트'와 두번째 '트'를 구분하지 못한다. 그래도 질의어의 글자 수가 많기 때문에 '트'외에 다른 글자를 이용하여 집합 기반 검색이 가능했었지만, 질의어의 글자 수가 적었다면 질의어의 중복 글자 비중이 높아지므로 정확도는 현저하게 떨어질 것이다.

질의어가 '뽀로로'처럼 글자 수가 적은 경우를 살펴보자. 기존 기법을 이용하면 원하는 POI 명칭이 상위에 랭크되지 않을 뿐더러 질의어 '뽀로'의 검색결과와 차이점도 없다. POI 데이터베이스에 '뽀로로'라는 레코드가 있다면 이 레코드 번호는 '뽀'와 '로' 글자를 포함한 레코드로 역 인덱스에 저장된다. 그래서 검색 시 '뽀로로' 뿐 아니라 '삿뽀로', '단뽀 가로수' 등과 같은 차수로 인식된다(차수는 검색결과에 질의어 문자의 포함 개수를 의미). 그러므로 질의어로 '뽀로로'같은 중복 글자를 포함한 단어가 입력될 경우 '뽀로로'보다 '삿뽀로'가 상위 랭킹되어 제일 먼저 출력된다. '뽀로로'나 '삿뽀로' 모두 질의어 문자인 '뽀'와 '로'를 포함하고 있어 어느 게 더 정확한지 알지 못하기 때문이다. (그림 3)은 기존 집합 기반 POI 검색 기법에서 중복 글자 질의어 처리과정을 보이고 있다.

(그림 3) 기존 집합 기반 POI 검색 기법의 중복 글자 질의어 처리과정



(Figure 3) Steps for handling a data including duplicate characters in existing set-based POI search algorithm.

중복 글자를 고려하지 못하는 문제는 글자 아이디 생성부터 텍스트 검색 단계까지 거의 모든 과정에서 발생한다. 특히 글자 아이디 생성 알고리즘, 역 인덱스 생성 알고리즘, 텍스트 검색 알고리즘에서 그 한계가 분명하다.

글자 아이디 생성 알고리즘에서는 글자의 유니코드로 해당 글자가 전체 POI 데이터베이스에 존재하는지만 확인한다. 한 레코드에서 최대 몇 개 포함하는지는 고려하지 않는다. 전체 POI 데이터베이스에서 하나의 레코드에 한 개만 존재하는 글자도 글자 아이디를 부여한다. 반면 전체 POI 데이터베이스에 여러 레코드에 존재하고 한 레코드 내에서 두 개 이상 존재하는 글자도 하나의 글자 아이디를 부여한다. 전체 POI 데이터베이스에 포함하는 개수, 한 레코드 내에 중복해서 포함하는 개수는 고려하지 않고 POI 데이터베이스에 해당 글자의 존재여부만 판단하여 존재하면 하나의 글자 아이디만을 부여한다.

역 인덱스 생성 알고리즘에서는 글자 아이디 생성 알고리즘에서 생성한 글자 아이디를 이용해서 역 인덱스를 생성한다. 역 인덱스 생성 과정에서도 한 레코드 내에서 해당 글자의 존재여부만 판단한다. 레코드 내에 몇 개를 포함하는지는 고려하지 않는다. 예를 들어 '로'를 한 개 포함한 '원효로'와 '로'를 두 개 포함한 '바로바로' 모두 '로'의 글자 아이디 하나에 모두 인덱싱된다. 그래서 레코드 내에서 몇 번 등장하는지 구분할 수 없는 문제가 생긴다.

텍스트 검색 알고리즘에서는 입력된 질의어를 읽은 후 역 인덱스를 이용하여 질의어를 포함하는 레코드를 검색한다. 역 인덱스에서 해당 글자가 한 레코드 내에서 몇 번 등장하는지 구분하지 못하고 존재여부만 알 수 있기 때문에 검색결과 역시 중복 글자를 고려하지 않은 결과가 출력된다. 입력된 질의어에 중복 글자가 있는 경우도 마찬가지이다.

기존 집합 기반 POI 검색 기법에서는 이렇게 한 레코드 내의 중복 글자를 고려하지 않는 한계가 있었다. 이에 본 논문에서는 이 같은 한계를 해결하기 위해 중복 글자를 구분하는 집합 기반 POI 검색 알고리즘을 제안한다.

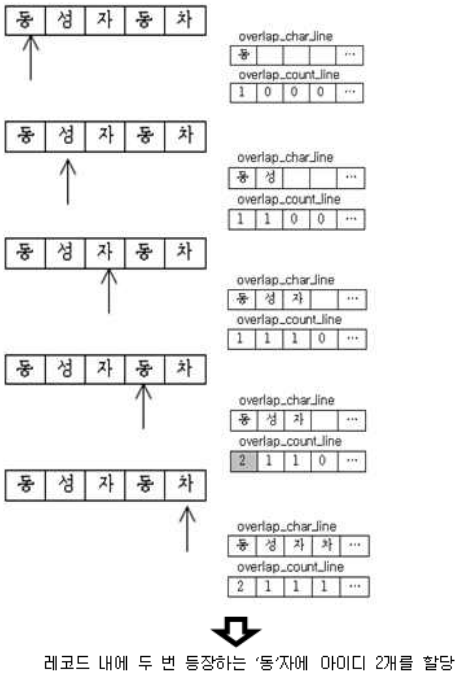
4. 중복 글자를 지원하는 집합 기반 POI 검색 기법

기존 검색 기법의 문제를 해결하기 위해 글자 아이디 생성 알고리즘, 역 인덱스 생성 알고리즘, 텍스트 검색 알고리즘 이렇게 세 부분을 개선하였다. 각 알고리즘의 개선 내용은 다음과 같다.

먼저 글자 아이디 생성단계에서 한 레코드 내의 중복 글자 여부를 식별한다. 이를 위해 한 레코드 내의 중복 글자 여부를 판단하는 1차원 배열 overlap_char_line, overlap_count_line과 전체 데이터베이스에서 1번이라도 레코드 내에

중복이 있는 글자를 모아 관리하는 1차원 배열 overlap_char, overlap_count를 생성한다. 1차원 배열 2개를 이용하여 각각 한 레코드에 포함된 글자와 그 글자의 중복 횟수를 레코드별로 기록하고 초기화한다. (그림 4)는 세부적인 동작과정을 보이고 있다.

(그림 4) 중복 글자를 지원하는 글자 아이디 생성 알고리즘



(Figure 4) Charater ID generating algorithm supporting classifying duplicate characters

앞 장에서 든 예시의 경우 그림 5와 같이 '바','로' 처럼 한 레코드 내의 중복이 있는 글자는 첫 번째로 등장하는 글자와 두 번째로 등장하는 글자를 구분하기 위해 2개의 아이디를 갖게 된다. (그림 5)는 보완된 집합 기반 POI 검색 기법에서 생성된 글자 아이디를 표현한 것이다. 3장의 (그림 3)에 나온 기존 집합 기반 POI 검색 기법에서 생성된 글자 아이디와 비교했을 때 '바', '로'가 1개씩 더 증가했다.

(그림 5) 중복 글자를 지원하는 글자 아이디

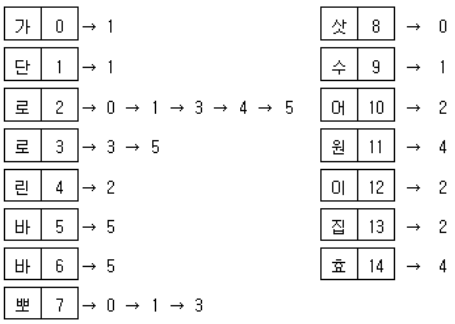
가	단	로	로	린	바	바	뽀	샷
0	1	2	3	4	5	6	7	8

(Figure 5) Charater ID supporting classifying duplicate characters

이렇게 레코드 내의 중복을 고려하여 생성된 글자 아이디는 역 인덱스 생성 알고리즘에 적용된다. 중복된 동일 글자를 구분해서 역 인덱스를 생성한다.

생성된 역 인덱스는 (그림 6)과 같다. '바'와 '로'가 두번 등장하는 레코드 번호를 저장했다. 기존 POI 검색 기법에서 생성된 역 인덱스와 비교해보면 늘어난 글자 아이디에 맞게 중복이 포함된 레코드가 추가되었음을 알 수 있다.

(그림 6) 중복 글자를 지원하는 역 인덱스

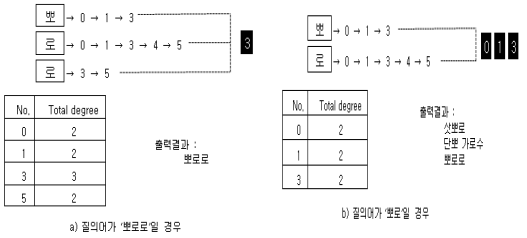


(Figure 6) Inverted indexes supporting classifying duplicate characters

마지막으로 텍스트 검색 알고리즘을 수정한다. 입력된 질의어 내의 중복 글자를 반영하여 전처리 작업을 해준다. 그 후 (그림 7)과 같이 블록 내 연산과 블록 간 연산에서 이 전처리 작업 결과에 맞추어 작업을 진행한다. 기존 POI 검색 기법과 같은 알고리즘으로 연산하지만 전처리 작업 결과에서 이미 질의어 내 중복 글자를 반영하고 있고, 역 인덱스에서 POI 데이터베이스 레코드 내 중복 글자를 반영하고 있기 때문에 레코드 내 중복 글자를 고려한 결과가 출력된다.

이렇게 보완된 집합 기반 POI 검색 기법은 기존에는 고려하지 못했던 레코드 내 중복 글자를 반영한다. (그림 7)은 보완된 알고리즘에 의해 검색을 수행하는 과정이다.

(그림 7) 질의어 내 중복 글자 처리를 지원하는 검색



(Figure 7) Searching for a data including duplicate characters

기존 집합 기반 POI 검색 기법에서는 '뽀로로'로 검색해도 '뽀로로' 외의 다른 데이터들이 상위에 랭킹되며 '뽀로로'와 '뽀로'를 구분하지 못해 검색결과가 동일했었는데, 이러한 문제점을 해결하였다.

4. 성능 평가

본 논문에서 제시한 개선된 알고리즘의 성능을 평가하기 위해 실험을 수행하였다. 중복글자가 있는 질의어 50개를 기존 집합 기반 POI 검색 기법과 개선된 집합 기반 POI 검색 기법에 입력하여 검색 결과를 비교하는 방식으로 진행했다.

질의어는 5자 미만에 두 번 이상 등장하는 글자를 적어도 1개 이상 포함한 단어로 무작위로 추출했다. 검색 성공 여부는 정답 데이터의 랭킹을 기준으로 판단했다. 상위 5위안에 들 경우 성공, 상위 20위 안에 들 경우 준성공, 그렇지 않은 경우에는 실패라 간주한다. 표 1은 실험결과를 정리한 것이다.

<표 1> 실험결과

구분	총	성공		준성공		실패	
기존 집합 기반 POI 검색 기법	50	11	22%	13	26%	26	52%
중복 글자를 지원하는 집합 기반 POI 검색 기법	50	40	80%	9	18%	1	2%

<Table 1> Test results

중복 글자를 포함한 질의어에 대해 실패율이 높은 기존 검색 기법에 비해 검색 성공률이 22%에서 80%로 향상된 것을 볼 수 있다. '동성자동차'의 경우, 기존 검색 기법에서는 '한성 자동차', '성원자동차' 등 정답과 거리가 먼 데이터들이 상위에 랭킹되었지만 보완된 검색 기법에서는 '동성자동차'가 1순위로 검색되었다.

수행한 실험을 통해 개선된 집합 기반 POI 검색 기법이 기존 집합 기반 POI 검색 기법보다 검색 성공률과 효율이 향상되었음을 알 수 있다.

6. 결론

기존 집합 기반 POI 검색 기법은 오질의어에 취약한 하드매칭 기법에 비해 정확한 검색결과를 출력하여 성능을 향상시켰다. 하지만 집합 개념을 기반으로 했기 때문에 동일 글자가 두 번 이상 등장하는 질의어나 레코드에 대한 처리가 미흡했다. 본 논문에서는 이런 문제를 해결하기 위해 중복 글자 처리를 지원하는 집합 기반 POI 검색 기법을 제안하였다.

제안된 집합 기반 POI 검색 기법에서는 한 레코드 내의 두 번이상 등장하는 중복 글자를 고려해서 글자 아이디를 부여하여 중복 글자를 인식할 수 있는 역 인덱스를 생성한다. 이 역 인덱스를 통해 중복 글자가 있는 질의어가 입력되어도 정확한 정보를 출력할 수 있고 이를 실험을 통해 확인하였다.

References

[1] Xinyan Zhu and Chunhui Zhou, "POI Inquiries and data update based on LBS," Proc. of the International Symposium on Information Engineering and Electronic Commerce(IEEC), pp. 730-734, 2009.

[2] Young-Kug Ham, Tae-Eun Kim,, "The Optimization path searching Method Development for Destination", Journal of Digital Contents Society , vol. 6, no. 1 pp. 55-62, 2005.

[3] Jin-Yong Moon, "Development of Spatial Object Converter for a Map Services in Mobile Environment", Journal of Digital Contents Society , vol. 13, no. 1, pp. 31-36, 2012.

[4] T. Tanaka, T. Uchihira, "Application of Mobile GIS Equipped with GPS to Field Survey with Public Participation," AIJ Journal of Technology and Design, vol. 14, no. 27, pp. 199-204, 2008.

[5] Hyang-Jin Lee, Jung-Hwa Choi, Young-Tack Park, "Semantic Point of Interest Detection from Large-scale GPS Data of Mobile Users", Journal of KIISE : Software and Applications, vol 39, no. 3, pp. 175-184, 2012.

[6] Eunbi Go, JaeWon Lee, Jongwoo Lee, An Efficient Set-based POI Search Algorithm [Online].Available: http://mm.sookmyung.ac.kr /~zinc/SetPOISearch.pdf

[7] AhYeon Jin, JaeWon Lee, Jongwoo Lee, "Measuring Method of String Similarity for POI Data Retrieval", Journal of KIISE : Computing Practices and Letters, vol. 19, no. 4, pp. 177 - 185, 2013.

[8] JaeWon Lee, "An Auto-corrective POI Retrieval Method based on Set-based Retrieval and Successive Matching Degree", Journal of KIISE : Computing Practices and Letters, vol. 19, no. 9, pp. 462 - 468, 2013.

[9] Yuen- Hsien Tsieng, Da-Wei Juang and Shiu- Han Chen, "Global and Local Term Expansion for Text Retrieval," Proc. of the Fourth NTCIR Workshop on Evaluation of Information Retrieval, 2004.

[10] Soonjin Kwon, Chuleui Hong, Wonil Kim,, "Revealing Hidden Relations between Query-Words for an Efficient Inducing User's Intention of an Information Search", Journal of the Institute of Electronics Engineers of Korea, vol. 49, no. 2, pp. 44 - 52, 2012.

[11] Ian Ruthven and Mounia Lalmas, "A survey on the use of relevance feedback for information access systems," Knowledge Engineering Review, vol. 18, no. 2, pp. 95 - 145, 2003.

고 은 별

2009 ~ 2012: 숙명여자대학교 멀티미디어과학과 학사
2013 ~ 현재 : 숙명여자대학교 멀티미디어과학과 석사

관심분야 : 검색시스템, 자연어처리, 알고리즘, 모바일 소프트웨어

이 종 우

1990년 : 서울대학교 컴퓨터공학과 (학사)
1992년 : 서울대학교 컴퓨터공학과 대학원(석사)
1996년 : 서울대학교 컴퓨터공학과 대학원(박사)
1996～1998년: 현대전자㈜ 정보시스템사업본부 과장
1998～1999년: 현대정보기술㈜ 책임연구원
1999～2002년: 한림대학교 정보통신공학부 조교수
2002～2003년: 광운대학교 컴퓨터공학부 조교수
2003～2004년: 아이닉스소프트㈜ 개발이사
2004～현재 : 숙명여자대학교 멀티미디어과학과 교수
2008년 : 뉴욕주립대 스토니브룩 Research Scholar
관심분야: Mobile System Software, Storage Systems, Computational Finance, Cluster Computing, Parallel and Distributed Operating Systems, Embedded System Software