http://dx.doi.org/10.7840/kics.2016.41.2.241

Open API 기술문서를 이용한 Open API 파라미터 유사도 비교

김 상 일◆, 김 화 성°

Similarity Comparison Among Open API Parameters Using Open API Description Document

Sang-il Kim◆, Hwa-sung Kim°

요 약

최근 스마트 디바이스의 보급으로 인해 스마트 디바이스 기반의 다양한 서비스가 창출되고 있으며, 특히 단순한 서비스 제공이 아닌 사용자의 상황, 환경에 맞는 사용자 맞춤형 서비스에 대한 수요가 증가하고 있다. 하지만 현재의 사용자 맞춤형 서비스는 불특정 다수 사용자를 위해 프로그램 제작자가 만드는 서비스이기 때문에 일반 사용자는 자신의 상황에 완벽하게 적합한 서비스를 제공받을 수 없는 문제점이 있다. 따라서 본 논문에서는 사용자 상황이 고려된 맞춤형 서비스 제공을 위한 자동 매쉬업의 요소기술로서, Open API 기술 문서에서 단어의 상호정보량을 산출하여 Open API 파라미터간의 의미적 유사도로 정의하고 비교하는 방안에 대해 연구 하였다.

Key Words : Open API, Amount of mutual information, Automatic mashup

ABSTRACT

The recent spread of smart devices has led to creating a variety of services based on the smart device, and the needs for the user-centric services that fit the individual users according to their situations and characteristics are increasing. However, current services can not fulfil the individual requirement of individual user, because these services are intended for unspecified individual. This paper, as a key technology of automatic user-centric service mash-up considering the situation of individual user, investigated the similarity comparison method between the Open API parameters by calculating the amount of mutual information of the parameters extracted from the Open API documents.

Ⅰ. 서 론

현재 스마트 디바이스를 통해 사용자들에게 제공되고 있는 서비스는 단순한 하나의 Open API 서비스뿐만 아니라 다양한 Open API를 결합한 매쉬업 서비스로 제공되고 있다. 매쉬업 서비스는 서비스를 만드는 개발자가 제공하고자하는 서비스의 목적에 맞게 다수의 Open API를 활용하여 프로그래밍을 통해 제작되고 있다. 하지만 개발자 중심의 매쉬업 서비스는 전문적인 개발자들이 불특정 다수 사용자들의 요구사항을 분석하고 요구사항에 맞는 서비스를 개발하기 때문에, 특정 사용자에게 맞춤화된 서비스를 제공하지 못하는 문제점이 있다.[1,2] 이러한 문제점을 해결하기 위해서 자동 매쉬업 기술에 대한 관련 연구가 활발히 진행 중에 있는데, 자동 매쉬업을 위해서는 Open API 파라미터 간의 합성 가능 여부를 판단하는 방안에 대한 연구가 필요하다. 따라서 본 논문에서는 Open API 기술 문서의 파라미터 이름에서 고유단어를 추출하여 단어 간의 상호 정보량을 기준으로 파라미터간의 유사도를 산출하여 매쉬업 가능 여부를 판단할 수 있는 Open API 파라미터 간의 유사도 비교 방안에 대해 연구하였다.

Ⅱ. 본 론

2.1 기존 유사도 비교 방안

보통 단어의 유사도 비교 방법은 검색 사이트에서 검색을 할 때나 단어를 정확하게 입력하지 못했을 경우 추천 검색어, 또는 비슷한 검색어를 제공하기 위해 이용된다. 두 단어 간의 유사도 측정의 기본적인 기법

※ 이 논문은 2014년도 광운대학교 교내 학술연구비 지원에 의해 연구되었음
※ 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신연구기반구축사업의 일환으로 수행하였음. [12221-14-1001, 차세대 네트워크·컴퓨팅 플랫폼연구 기반구축]
◆ First Author : Department of Electronics and Communications Engineering, Kwangwoon University, rlatkd234@kw.ac.kr, 학생회원
° Corresponding Author : Department of Electronics and Communications Engineering, Kwangwoon University, hwkim@kw.ac.kr, 종신회원
논문번호 : KICS2016-01-011, Received January 18, 2016; Revised February 16, 2016; Accepted February 16, 2016

은 단어를 잘게 자르고 난 다음에 해당 단어의 조각들을 각각 비교 하여 같은 조각의 개수를 산출하여 해당 단어들 간의 유사도를 판단하는 방식을 사용한다. 이와 같은 방법을 bi-gram을 이용한 유사도 비교 방법이라고 한다. 하지만 bi-gram 방식을 통한 유사도 비교 방법은 단어의 의미를 고려하지 않고 단순하게 구조가 비슷한 단어를 찾기 때문에 파라미터의 합성에 적용하기에는 한계가 존재한다.

2.2 제안하는 파라미터 유사도 비교방안

현재 Open API간의 매쉬업은 개발자가 Open API 파라미터에 관련된 정보를 기술하고 있는 Open API 기술문서를 참조하여 파라미터 간의 유사도를 기반으로 합성하고자 하는 Open API간의 합성 가능 여부를 판단한다. 합성이 가능한 것으로 판단된 경우, 하나의 Open API에 요구 되는 값을 입력하고, Open API가 산출하는 결과 값을 파싱을 통해 인터페이스 또는 다른 기능을 하는 Open API의 입력 값으로 제공하는 형태로 매쉬업이 이루어지고 있다. 하지만 기존의 개발자가 제작하는 매쉬업 서비스가 아닌 자동화된 매쉬업 프레임워크에서는 사용자가 Open API 간의 매쉬업 가능 여부를 판단하는 것이 아니라, 컴퓨터가 Open API 간의 매쉬업 가능 여부를 판단할 수 있어야 한다. 본 논문에서 제안하는 파라미터 유사도 비교 방안 역시 Open API 기술문서를 사용하였다. Open API 기술문서에는 Open API에 대한 설명뿐만 아니라 Open API에서 제공하는 각 파라미터에 대한 상세 정보도 제공되기 때문에 Open API 간의 합성 가능 여부 판단을 위한 기준인 파라미터 간의 유사도 산출에 충분히 사용 가능하다.

본 논문에서는 파라미터 이름을 사용하여 파라미터간의 유사도를 산출한다. 파라미터 이름은 'geo', 'user'와 같이 의미를 나타내는 단어만 사용하는 경우, 'geo_id', 'user_name' 과 같이 의미를 나타내는 단어와 합께 타입을 나타내는 단어가 함께 사용되는 경우 'geo_location'과 같이 의미를 나타내는 단어가 두 개 이상으로 구성된 파라미터 이름이 존재한다. 따라서 제안하는 알고리즘에서는 파라미터 이름에서 의미를 나타내는 단어와 타입을 나타내는 단어를 분리하여 단어 간의 유사도를 산출한다. 하지만, 'geo_id'와 'location_name'의 경우 의미를 나타내는 단어가 비슷하기 때문에 합성 가능한 것처럼 보이지만 파라미터 타입이 다르기 때문에 합성 될 수 없다. 이러한 이유로 제안 하는 알고리즘에서는 의미를 나타내는 단어가 유사하더라도 타입이 같지 않으면 두 파라미터는 유사하지 않다고 판단한다. 또한 'user'와 'user_id'와 같이 타입이 포합된 단어와 타입이 포함되지 않은 단어의 경우도 두 파라미터는 유사하지 않다고 판단한다. 즉, 'user_id', 'person_id'와 같이 의미 유사도가 높은 단어와 같은 타입의 단어로 구성된 파라미터만을 유사한 파라미터로 정의한다. 각 파라미터 이름에서 추출한 의미를 나타내는 단어들의 경우 두 단어 간의 유사도를 상호정보량 공식인 식(1)과 같이 산출하는데 이를 파라미터간의 유사도로 정의한다. 하지만 타입을 나타내는 단어의 경우, 상호정보량은 산출하지 않는다.

$$I(w_1,w_2) = \log \frac{P(w_1,w_2)}{P(w_1)P(w_2)} \quad (1)$$

식 (1)에서 *P(w)*는 Open API의 모든 파라미터 이름들 중에 특정단어 w가 포함된 파라미터가 존재할 확률이다. 두 단어 w1과 w2에 대해 *P(w1)*P(w2)*와 비교하여 Open API의 모든 파라미터 이름들 중에 두 단어 *w1,w2*가 함께 포함된 파라미터가 존재할 확률 *P(w1, w2)*가 높을수록 두 단어는 상호 정보량이 크다고 정의한다. 즉, 상호정보량이 높은 두 단어가 서로 다른 파라미터 이름에 포함되어 있으면 두 파라미터는 의미적으로 관계성이 높다고 할 수 있다. 따라서 두 개의 파라미터 이름에서 획득한 각 단어 간의 상호정보량은 각 단어들이 포함된 두 개의 파라미터들 간의 유사도로 정의 할 수 있다.

그림 1은 상호정보량을 기반으로 파라미터 간의 유사 여부를 판단하는 시나리오를 나타낸다. 먼저 그림 1에 보인 것과 같이 Open API 기술 문서 내에서 파라미터 이름을 추출하고 해당 파라미터 이름에서 의미를 나타내는 단어를 도출하여 해당 단어가 포함된 파라미터의 개수를 출현 빈도 테이블로 생성한다. 이후 두 단어가 함께 나오는 파라미터의 개수를 동시 출현 빈도 테이블로 생성한다. 두 단어들 간의 상호 정보량은 두 단어의 출현 빈도수와 동시 출현 빈도수를 기반으로, 식 (1)에 의해 산출한다. 이후 상호정보량을 바탕으로 두 파라미터들의 유사 여부를 판단할 수 있다. 본 논문에서는 상호정보량을 파라미터 유사도로 정의하였기 때문에 상호 정보량에 임계값을 적용하여 임계값 기반의 단어 유사도 그래프를 생성한다. 임계값이란 유사 단어들 간의 군집 형성을 위한 최소값으로, 임계값 이내의 값을 갖는 단어들은 같은 군집으로 표현된다. 같은 군집에 속한 단어들은 서로 상당히 유사한 의미를 갖는다는 것을 뜻한다. 즉, 비교하고자 하는 두 파라미터 이름에서 도출한 의미를 나타내는 단

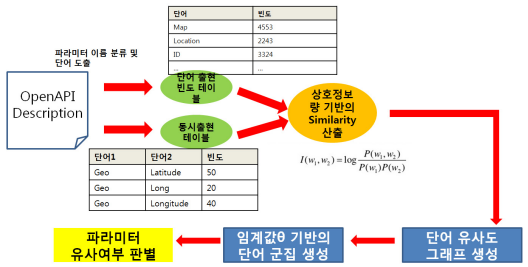

그림 1. 상호정보량 기반의 파라미터 유사 여부 판별 과정
Fig. 1. Similarity Decision Procedure between Parameters based on Amount of Mutual Information

어가 같은 그래프에 포함되고, 타입을 나타내는 단어가 같은 경우 두 파라미터는 유사하다고 판단한다.

Ⅲ. 실험 결과

그림 2는 제안하는 방안의 상호정보량 산출 결과다. 본 논문에서는 제안 방법의 성능 실험을 위해 Programmableweb 사이트에서 약 850개의 Open API 기술문서를 획득하여 사용하였다.

표 1은 기존 알고리즘과 제안하는 알고리즘의 유사 여부 판정 결과의 정확도를 나타낸 표이다. 본 논문에서는 유사 여부 판정 정확도의 확인을 위해 Programmableweb[3]에서 획득한 850 개의 Open API 기술문서 중에서 실제로 매쉬업에 사용된 300개의 파라미터들을 정답지로 사용하였다. 300개의 파라미터들을 정답지로 사용한 이유는 실제로 매쉬업에 사용된 300개의 파라미터들은 이미 서비스로 제공되고 있기 때문에 매쉬업이 가능하다고 볼 수 있기 때문이다. 표 1에 보인 정분류율은 정답지를 기반으로 제안 방법이 해당 파라미터간의 유사 여부 판정을 정답지에 맞게 정확하게 판단한 비율을, 오분류율은 그렇지 못한 경우를 의미한다. 표 1에서 보는 것과 같이 기존의 bi-gram을 사용한 알고리즘은 54%의 낮은 성능을 보이지만 제안하는 알고리즘은 75.4%의 높은 성능을 보이는 것을 알 수 있었다. 기존의 알고리즘인 bi-gram 알고리즘은 단순히 단어의 구조만을 비교하기 때문에 약어 또는 같은 회사의 Open API 파라미터는 비교적

Word 1	Word 2	Sim(MI : 상호 정보량)	Sim(상호정보량 정규화)
company	stock	5.32	2.13217
stock	market	6.54	2.42102
stock	phone	1.02	4.236659
movie	star	6.59	0.988732

그림 2. 제안하는 알고리즘의 상호 정보량 산출 결과
Fig. 2. Mutual Information Calculation Results of the Proposed Algorithm

표 1. 기존 및 제안알고리즘의 파라미터 유사도 산출 정확도
Table. 1. Accuracy Rates of Parameter Similarity Calculation of Existing & Proposed Algorithms

	기존 알고리즘(bi-gram)의 유사도 산출정확도	제안하는 알고리즘의 유사도 산출 정확도
유사도 산출 정확도	54%	75.4%

정확하게 판단하지만, 의미가 같은 다른 회사의 Open API 파라미터는 정확히 판단하지 못하는 문제점과 파라미터에서 사용되는 같은 의미의 다른 단어, 예를 들어 'Geo', 'Location' 과 같은 단어는 유사한 것으로 판단하지 못하는 문제점이 있다. 이와 달리 본 논문에서 제안하는 방법은 기존 알고리즘과 달리 높은 정확도를 보이는 것을 실험을 통해 보였다. 하지만 본 논문에서 제안하는 방법은 빈도수에 의해 상호정보량의 값이 결정되기 때문에 의미 없이 반복되는 단어에 의해 그 정확도가 낮아 질 수 있는 문제점이 존재한다.

Ⅳ. 결 론

본 논문에서는 자동화된 매쉬업 프레임워크를 개발하기 위한 요소 기술로써 상호정보량 기반의 파라미터 유사도 비교를 통한 파라미터 유사 여부 판단 방안에 대해 연구하였다. 제안하는 알고리즘의 성능은 기존의 bi-gram기반의 Open API 파라미터 유사도 판별 방안 보다 더 높은 성능을 보이는 것을 실험을 통해 확인 하였다. 하지만 상호정보량 기반의 분류는 빈도수에 의해 상호정보량의 값이 결정되기 때문에 의미 없이 반복되는 단어에 의해 그 정확도가 낮아 질 수 있는 문제점이 존재한다. 따라서 향후 연구로는 더 높은 정확도를 제공하기 위해 상호정보량 도출 방안을 수정한 Open API 파라미터 비교 방안과 매쉬업 방안에 대해 연구를 진행 할 예정이다.

References

[1] A. K. Moon, Y. M. Park, and S. G. Kim, "Technical trends of semantic annotation for semantic web services," *Electronics and Telecommun. Trends*, vol. 25, no. 2, pp. 121-131, Apr. 2010.

[2] Y. J. Lee and J. H. Kim, "Semantically enabled data mashups using ontology learning method for web APIs," in *Proc. 2012 Computing, Commun. Appl. Conf.*, pp. 304-309, Hong Kong, Jan. 2012.

[3] Programmableweb, http://www.programmableweb.com