Journal of Digital Contents Society Vol. 12 No. 4 Dec. 2011(pp.429 - 436)

시맨틱 주석을 이용한 내용 기반 데이터 검색

김병곤*, 오성균**

요 약

인터넷검색의 대상이 되는 각종 문서, 이미지, 동영상 등의 자료가 늘어날수록 이에 대한 효율적인 검색의 문제가 중요시되고 있다. 효율적인 검색의 관점은 초기의 키워드 중심의 검색에서 자료가 지니는 의미적인 요소들을 종합적으로 판단하여 이들의 연관성을 찾아 검색하는 의미적 검색의 방향으로 진행되고 있다. 이에 따라, 각종 자료에 대한 의미적 검색을 위하여 메타데이터 처리를 위한 시맨틱 주석을 생성, 운영하는 시스템들이 연구되어 왔다. 그러나, 동일한 종류의 자료에 대한 주석 위주로 진행되었고, 각기 다른 방법과 형태로 생성된 주석 데이터 간에는 호환적인 검색이나 처리가 어렵다. 본 연구에서는 이 문제를 해결하기 위하여 다양한 주석문서를 내용분석에 따라 단계별 형태로 분류하고, 상이한 종류의 자료 간에도 검색이 가능하도록 문서간의 유사도를 측정하는 방법을 제시하였다. 주석문서간의 유사도 측정은 소스문서와 유사도가 높은 주석문서를 검색하여 결과적으로 자료의 종류나 형태에 상관없이 가장 유사한 내용을 지니는 문서나 이미지, 동영상 등을 검색하는데 사용할 수 있다.

Content based data search using semantic annotation

Byung-Gon Kim*, Sung-Kyun Oh**

Abstract

Various documents, images, videos and other materials on the web has been increasing rapidly. Efficient search of those things has become an important topic. From keyword-based search, internet search has been transformed to semantic search which finds the implications and the relations between data elements. Many annotation processing systems manipulating the metadata for semantic search have been proposed. However, annotation data generated by different methods and forms are difficult to process integrated search between those systems. In this study, in order to resolve this problem, we categorized levels of many annotation documents, and we proposed the method to measure the similarity between the annotation documents. Similarity measure between annotation documents can be used for searching similar or related documents, images, and videos regardless of the forms of the source data.

Keywords : Semantic annotation, RDF/XML, Similarlity, Schema

1. 서론

급속도로 진행되는 웹의 보편화로 인하여 데이터의 양이 급격히 증가하고, 디지털카메라, 스마트폰, 테블릿PC 등의 휴대용 무선 디바이스의 고성능화와 모바일 환경의 개선에 따라 이들 기기에서 생산되는 콘텐츠의 양이 더욱 빠르게 증가하고 있다[1]. 따라서 텍스트 정보에 대한 키워드 검색뿐 만 아니라 멀티미디어 데이터를 검색 질의의 대상으로 하는 내용 기반 데이터 검색을 지원하는 검색 시스템의 개발이 활발히 진행되고 있다. 현재까지 진행되고 있는 내용기반 검색의 가장 일반적인 형태는 대상 문서나 멀티미디어 데이터에 관한 키워드 형태의 메타데이터 태그(Tag)를 관리자나 사용자가 부여하고 인

※ 제일저자(First Author) : 김병곤
접수일:2011년 10월 21일, 수정일:2011년 11월 26일
완료일:2011년 12월 04일
* 부천대학 e-비즈니스과
bgkim@bc.ac.kr
** 서일대학 소프트웨어과
▣ 본 논문은 2010년도 서일대학 학술연구비에 의해 연구되었음.

덱스를 구축하여 검색하는 방법이었다. MP3화일의 ID3태그, 디지털 카메라 JPEG 이미지에 첨부되는 EXIF 메타데이터, 이미지에 첨부되는 여러 가지 비형식의 데이터들이 쉽게 접할 수 있는 메타데이터의 예이다. 그러나 태그 방식의 내용 기반 검색은 복잡하고 다양한 데이터의 연관 검색을 수행하는데 많은 한계를 지니고 있다. 이러한 한계를 극복하기 위하여 좀 더 구조화된 메타데이터의 형태인 주석(Annotation)과 시맨틱웹의 개념을 이용하여 대량의 문서, 이미지, 각종 멀티미디어 데이터를 보다 효율적으로 저장, 검색하기 위한 방법으로 시맨틱 주석(Semantic annotation)이 연구되고 있다[1,2].

여러 다양한 형태로 구성되어 있는 주석의 형태를 시맨틱웹 기술을 이용하면 좀 더 효율적으로 표현하고 활용할 수 있다. 이는 시맨틱웹 기술이 주석으로 표현되어 있는 웹 자원들을 분류하고 서로 연관지을 수 있는 기반기술을 제공하기 때문이다. 시맨틱웹을 구현하는데 핵심적인 역할을 하는 것이 온톨로지이다. 비형식적이고 주관적인 표현의 웹상의 단어들을 컴퓨터가 처리 가능한 형식으로 변환이 가능하도록 해주는 것이 온톨로지의 개념이다. 온톨로지는 특정 도메인의 공유할 수 있는 개념, 관계와 제약사항 등을 나타내는 공통의 단어들을 제공한다. 분산되어 있는 시스템들이 온톨로지를 기준으로 데이터를 주고 받고 상대방의 데이터를 자동으로 처리가 가능해진다. 그러므로, 멀티미디어 자료나 웹문서에 대한 고차원의 검색을 위하여 시맨틱웹의 개념을 이용하기 위하여 각 시스템들은 여러 메타데이터 표준을 기반으로 각 시스템에 적합한 시맨틱 주석을 생성하여 사용하고 있다.

지금까지 더블린코어, VRA, MPEG-7과 같은 메타데이터 표준에서 주석에 대한 표현과 생성에 대한 중요한 기능들을 제시하였고, 이와 관련하여 MnM[3], SHOE[4], KIM[5]과 같은 여러 온톨로지 주석 툴 시스템을 통하여 시맨틱 주석에 대한 많은 시도들이 계속되어 왔다. 언급된 초기의 연구들은 HTML과 같은 인터넷 상의 문서 자료의 주석에 대한 연구에서 시작하였으며 점차 이미지, 동영상과 같은 멀티미디어 데이터에 대한 주석으로 연구의 범위를 넓혀가고 있다. 각 시스템들은 동일한 종류의 자료에 대한 주석, 즉 html 문서에 대한 주석을 다루는 시스템, 디지털 이미지에 대한 주석을 다루는 시스템, 혹은 동영상 자료에 대한 주석을 다루는 시스템과 같이 각기 다른 방법과 형태로 연구되었으나, 이들 시스템들로부터 생성된 주석 데이터 간에는 검색이나 처리가 어렵고, 서로 다른 종류의 주석문서간의 검색을 위해서는 유사도 측정과 같은 메카니즘을 필요로 한다.

본 연구에서는 이 문제를 해결하기 위하여 다양한 주석 문서를 RDF/XML형태로 변형하고, 변형된 주석 데이터를 내용에 따라 단계별로 분류하는 방법과 이를 바탕으로 한 유사도 측정 방법을 제시한다. 주석문서간의 유사도 측정은 소스문서와 유사도가 높은 주석 문서를 검색하여 결과적으로 가장 유사한 내용을 지니는 문서나 이미지, 동영상 등을 검색하는데 사용할 수 있다. 먼저, 다양한 형태의 주석 데이터를 3가지 형태로 단계별로 분류하며, 주석 간의 유사도를 측정할 때 단계, 스키마, 요소, 속성을 가지고 유사도를 측정할 수 있고, 이를 통하여 검색할 때에도 검색하고자 하는 의도에 따른 검색이 가능해 진다.

2. 주석의 표현

시맨틱 주석은 인스턴스 데이터에 대한 메타데이터를 생성하고 이를 온톨로지 클래스와의 매핑을 통하여 새로운 정보를 제공하는 것을 목표로 한다. 문서, 이미지, 동영상 등의 인스턴스 데이터에 주석을 부여할 때 다음과 같은 사항들을 고려하여 작성하게 된다. 주석은 수동 혹은 자동으로 생성한다. 일반적으로 주석의 수동 생성은 가장 적당한 레벨의 추상화된 정확한 데이터를 추출하는데 효율적이다. 하지만, 이는 많은 시간과 비용을 요구한다. 또한, 주석 작업자의 주관적인 경향에 따라 동일한 대상에 대한 전혀 다른 주석이 생성될 수 있다. 반면에, 자동 특성 추출 방법에 의한 주석 방법은 상대적으로 빠르고 저렴하며 시스템화 하기가 용이하지만 대부분의 자동 주석 추출 방법은 많은 응용분야에 있어서 아주 낮은 단계의 추상화 데이터만을 생성하는 경우가 많다. 그러므로 낮은 단계로 추상화된 주석 데이터는 실제 응용 분야에서 필요로 하는 고급 단계의 추상화 주석 데이터와의 차이

인 의미적 격차(Semantic gap)를 줄이는 주석의 생성을 목표로 하여야 한다.

주석을 생성할 때 일반화된 주석과 특성화된 주석을 구분할 수 있다. 자료의 주석을 생성할 때 특정한 목적이나 업무를 고려치 않고 일반적인 내용으로 주석을 생성하면 비용 측면에서 비효율적일 수 있다. 응용 분야가 결정된 후 이미 생성된 주석의 내용이 응용분야와 맞지 않거나 상이한 추상화 방법을 사용하는 경우에 다시 주석을 생성해야 하기 때문이다. 반면에, 특정업무에 특성화된 주석은 다른 응용 분야의 업무의 주석으로 사용하기에는 너무 세부적이어서 재사용하기에 어려운 점들이 존재한다. 그러므로, 이러한 두가지 상반 관계를 모두 고려하여 생성할 주석을 미리 예측하여 생성하여야 한다.

주석 생성 작업을 수행할 때 파악해야 할 다양한 형태의 메타데이터를 분류하면 문서나 이미지 자체 속성에 대한 서술과 서술 대상인 개체, 사람, 개념 등의 속성에 관한 서술로 분류할 수 있다. 첫 번째로 데이터 자체에 대한 서술은 문서의 제목, 작성자, 이미지의 해상도, 형식, 생성날짜 등이 해당한다. 두 번째 분류는 서술대상에 대한 묘사이며, 객관적인 관찰에 대한 결과를 서술하거나 주관적인 해석에 의한 결과를 서술하는 경우가 있을 수 있다.

위에서 언급된 사항들을 고려하여 주석 문서를 생성할 때 마지막으로 주석에서 사용할 메타데이터 표준 용어(Vocabularies)의 집합과 주석 문서의 형태를 결정하여야 한다. 주석 문서의 서술을 위하여 다양한 용어 집합이 필요하며, 과학이나 스포츠 등과 같은 특정 도메인에 관한 용어에서부터 도메인과 상관없이 일반적으로 통용되는 용어를 사용할 수도 있다. 대부분의 응용 프로그램에서는 이러한 특성들을 서술하기 위해 일반적으로 대중화 되어있는 더블린코어, VRA, MPEG-7 등의 메타데이터 표준 용어 집합을 사용한다. 더블린코어는 메타 데이터들에서 사용되는 기초적인 내용들을 표준화하여 검색 및 처리가 용이하도록 15개의 프로퍼티로 구성된 표준화된 메타 데이터 요소 집합이다[6]. 동영상, 소리, 이미지, 텍스트, 웹 페이지 등의 디지털 매체들을 기술하는 데 널리 사용되며, 보통 XML과 RDF를 사용하여 구현된다. 비디오, 오디오, 이미지, 텍스트뿐만 아니라 웹페이지와 같은 복합 매체에도 쉽게 적용이 가능하다. 단순한 구조의 시스템 개발, 교환 및 통합용 메타데이터, 응용 프로파일 개발의 기본 메타데이터 등 매우 많은 분야에 사용이 가능하다. VRA코어 카테고리는 예술, 건축, 대중 문화 및 민속 문화 유물, 예술 작품 등의 저작물에 대한 시각적 자료를 기술하기 위한 여러 가지 메타데이터 요소를 제공한다[7]. 멀티미디어 콘텐츠를 기술하는 표준인 MPEG-7은 XML 스키마를 사용하여 정의할 수 있다. 멀티미디어로 구성된 데이터베이스에서 정보를 쉽게 탐색하여 추출할 수 있도록, 표준화된 멀티미디어 정보 표현 방식을 제공하기 위한 국제 표준이다[8,9].

주석 생성시에는 위에서 언급된 웹 데이터의 프로퍼티나 내용을 표현하기 위하여 이미 발표된 여러 가지 표준의 요소와 속성과 시스템 자체적으로 구성한 스키마의 요소를 사용하여 주석을 생성할 수 있다. 예를 들어, 더블린코어나 VRA를 이용하여 기본적인 사항을 서술하며, 문서나 이미지에 표현되어 있는 내용은 상당히 다양하고 넓은 분야에 걸쳐 있기 때문에 내용적인 측면에서의 서술을 위하여 이미 존재하는 여러 다양한 용어를 사용하거나 새로운 특정 영역의 스키마를 만들어서 사용할 수도 있다.

시맨틱 주석 시스템은 시맨틱 주석을 이용하여 정보검색을 수행하기 위한 온톨로지, 지식베이스관리, API 접근 방법, 저장 구조, 사용자 인터페이스 등을 종합적으로 제공한다. 시스템에 따라 이중 일부의 기능만을 제공하는 경우도 있다. MnM[3]은 자연어 처리 방식으로 주석을 생성한다. 순차적으로 입력되는 단어들을 사용하여 규칙을 생성하고 생성된 규칙들을 가지고 텍스트에 시맨틱 태그를 삽입하며, 교정규칙에 의하여 좀 더 정확한 주석을 부여하는 방식을 사용한다. KIM[5]은 시맨틱 주석, 인덱싱, 검색을 위한 서비스와 하부 구조를 제공하는 시스템이다. 온톨로지와 지식 베이스를 이용하여 정보 추출을 수행하며, 하부 저장 구조로서 RDF를 위한 SESAME를 사용한다. 온톨로지 KIMO는 기본적인 엔티티 클래스와 관계, 제약 사항 등을 표현하며 이를 바탕으로 시맨틱 주석을 표현한다. M-OntoMat-Annotizer[10]는 이미지나 비디오와 같은 멀티미디어 자료를 분석하여 시맨틱 주석을 생성하는 시스템이다. MPEG-7 비쥬얼 설명

자(Descriptions)와 온톨로지를 연관하여 주석을 표현 할 수 있도록 하였다. 핵심 온톨로지로는 DOLCE를 사용하며 기본적인 표현 언어는 RDFS를 사용하였다.

주석문서 생성에서 중요한 결정 사항은 주석 문서의 형식을 결정하는 것이다. 위에서 언급된 시스템들을 참고해보면 일반적으로 상호 호환성을 가장 잘 보장하는 방법은 RDF/XML 문서 형태로 표현하는 것이다[11]. 아래 문서는 RDF/XML 문서 형태로 표현된 이미지에 대한 주석 문서의 일부이다.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:event="http://www.example.org/ontologies/event#"
  xmlns:locat="http://www.example.org/ontologies/location#"
  xmlns:lsc="http://www.example.org/ontologies/landscape#"
  xmlns:vra="http://e-culture.multimedian.nl/ns/vracore3.rdfs#">
    <dc:type rdf:resource="http://purl.org/dc/dcmitype/Image"/>
    <dc:description>Photo of Katerina Tzouvara</dc:description>
    <dc:creator>
      <foaf:Person>
        <foaf:familyname>Stabenaou</foaf:familyname>
        <foaf:firstname>Arne</foaf:firstname>
      </foaf:Person>
    </dc:creator>
    <dc:date>2002-12-04</dc:date>
    <dc:format>JPEG</dc:format>
    <vra:measurements.resolution>300 x 225px
    </vra:measurements.resolution>
    <locat:located_in_Continent>
      <locat:Continent>Asia</locat:Continent>
    </locat:located_in_Continent>
    <locat:located_in_Country>
      <locat:Country>Thailand</locat:Country>
    </locat:located_in_Country>
    <locat:located_in_City>
      <locat:City>
        <foaf:name>Phi Phi</foaf:name>
        <locat:belongs_to_Country>
          <locat:Country>Thailand</locat:Country>
        </locat:belongs_to_Country>
      </locat:City>
    </locat:located_in_City>
    <foaf:depicts>
      <lsc:Beach/>
    </foaf:depicts>
    <foaf:depicts>
      <lsc:Palm_Tree>Phoenix Dactyliphera</lsc:Palm_Tree>
    </foaf:depicts>
    <foaf:depicts rdf:resource=
       "http://www.example.org/ontologies/landscape#Sand"/>
    <foaf:depicts rdf:resource=
       "http://www.example.org/ontologies/event#Vacations"/>
    <foaf:depicts>
      <foaf:Person>
        <foaf:familyname>Tzouvara</foaf:familyname>
        <foaf:firstname>Katerina</foaf:firstname>
      </foaf:Person>
    </foaf:depicts>
  </rdf:Description>
</rdf:RDF>
```

문서의 앞부분은 주석에서 사용된 스키마 즉 용어 집합에 대한 네임스페이스를 정의한 부분이며, 하나의 주석 문서에서 여러 개의 스키마를 사용하여 데이터를 표현할 수 있다. 주석 문서의 내용을 살펴보면 dc네임스페이스를 사용하여 그림의 작가에 대한 기본적인 내용을 표현하였고, locat네임스페이스를 사용하여 그림에 나타난 장소, 사람에 대한 정보들을 구체적으로 표현하고 있다. 특히, foaf네임스페이스를 사용하여 그림의 내용에 대한 표현을 하고 있다.

RDF/XML 문서 형태로 표현된 주석의 구성을 분석해 보면 스키마, 요소, 속성, 값으로 구분할 수 있다. 스키마, 요소, 속성의 내용에 해당하는 값으로 표현된 두 개의 주석 문서에서 동일한 값의 데이터라도 주석 문서에서 사용된 스키마, 요소, 속성에 따라 서로 다른 내용일수도 있고, 같은 내용일수도 있다. 이러한 내용을 구분하여 주석 문서간의 유사도를 계산하고 이를 실제 인스턴스 데이터간의 유사도를 측정하는데 사용할 수 있다. 다음 장에서는 내용 기반 데이터 검색을 수행하기 위한 시맨틱 주석 문서간의 유사도 측정을 위하여 필요한 수식과 알고리즘을 제시하였다.

3. 주석문서 유사도 측정

주석 문서간의 유사도 측정은 웹상의 텍스트, 이미지, 동영상 등의 자료들을 자료의 형태와 상관없이 유사한 자료를 검색하는데 사용한다. 예를 들어, 다빈치의 모나리자 그림에 관한 텍스트 문서와 모나리자 이미지 파일을 유사한 데이터로 검색할 수 있는 기반을 마련할 수 있다. 각기 다른 시스템에서 생성된 주석 문서는 각기 다른 내용과 형태를 가지게 된다. 이러한 다양한 내용과 형태의 문서들을 한 시스템처럼 검색하기 위해서 다음과 같은 중간 단계를 거친다. 주석문서를 메타데이터 표현 방법 중에서 가장 일반적으로 많이 사용되고 있는 XML기반의 RDF/XML

형태로 표현하고, 다양한 형태의 주석 데이터를 다음과 같은 3가지 형태로 단계별로 분류한다. 단계별 분류의 목적은 주석 문서가 지니는 메타 데이터의 종류를 세분화하여 유사도 측정시에 좀 더 세밀한 비교가 가능하도록 하는데 있다.

<표 1> 주석 데이터 분류

분류	내용
물리적 주석 데이터	- 데이터의 물리적 저장형태, URL <dc:format>JPEG</dc:format>
논리적 주석 데이터	- 문서 , 이미지, 동영상 등의 제목, 작성자, 해상도, 생성날짜 등 <dc:creator> <foaf:Person> <foaf:familyname>Stabenaou </foaf:familyname> <foaf:firstname>Arne</foaf:firstname> </foaf:Person> </dc:creator> <vra:measurements.resolution>300 x 225px </vra:measurements.resolution>
의미적 주석 데이터	- 문서, 이미지, 동영상 등의 내용에 대한 서술이며, 객관적인 관찰에 대한 결과를 서술하거나 주관적인 해석에 의한 결과를 서술하는 경우 <foaf:depicts> <foaf:Person> <foaf:familyname>Tzouvara </foaf:familyname> <foaf:firstname>Katerina</foaf:firstname> </foaf:Person> </foaf:depicts>

단계별로 분류한 주석데이터는 주석문서간의 유사도를 측정할 때 단계별로 유사도를 측정할 수 있고, 검색할 때에도 검색하고자 하는 의도에 따라 단계별로 자세한 검색이 가능해 진다. 검색 의도에 따라 물리적 데이터, 논리적 데이터, 의미적 데이터에 대하여 각각 유사도를 측정할 수 있다.

다음으로 주석의 도메인, 속성, 값 일치도를 측정하여 유사도 측정에 사용할 수 있다.

<표 2> 유사도 관계 측정 항목

분류	내용
스키마	주석을 표현하는데 사용된 스키마
요소	주석을 표현하는데 사용된 스키마 요소를 의미
속성	주석을 표현하는데 사용된 요소의 속성을 의미
값	주석을 표현하는데 사용된 요소의 실제 값을 의미
단계	주석데이터의 3가지 단계

<표2>에서 언급한 바와 같이 유사도 측정을 위하여 크게 스키마, 요소, 속성, 값, 단계의 5가지 측면에서 주석 간의 유사도를 측정하도록 한다. 가장 유사도가 높은 경우는 동일 스키마의 요소와 속성을 사용하여 표현되고 동일한 단계의 일치하는 값을 지니는 요소가 많은 경우로 간주한다. 두 번째는 스키마, 요소, 속성, 단계의 일부만이 일치하면서 동일한 값을 지니는 경우이고, 다음으로는 스키마나 요소, 속성, 단계가 일치하지 않더라도 동일한 값을 지니는 요소가 많은 경우에도 가중치를 조정하여 유사도를 부여한다.

본 연구에서는 식(1)에서와 같이 주석 간의 유사도를 측정하기 위한 함수를 제시한다.

$$S(A_a,A_b) = a_1 \cdot S_{exactmatch}(A_a,A_b) + a_2 \cdot S_{partiallymatch}(A_a,A_b) + a_3 \cdot S_{valuematch}(A_a,A_b)$$
$$단,\ a_1 > a_2 > a_3 > 0 \quad (1)$$

유사도 함수 S는 <표2>의 유사도 관계 측정에서 제시한 항목에 따라 두개의 주석간의 유사도를 수치 값으로 표현하였다. $S(A_a,A_b)$는 주석 A_a와 A_b간의 유사도를 나타낸다.

$S_{exactmatch}(A_a,A_b)$는 두 개의 주석 중에 값이 일치하는 요소중에서 스키마, 요소, 속성, 단계가 모두 일치하는 요소의 개수를 의미한다. $S_{partiallymatch}(A_a,A_b)$는 두 개의 주석 중에 값이 일치하는 요소 중에서 스키마의 동일 요소, 속성, 단계가 일부 일치하는 요소의 개수를 의미한다. $S_{valuematch}(A_a,A_b)$는 두 개의 주석 중에 스키마, 요소, 속성, 단계와 일치하지 않고 값만이 일치하는 요소의 개수를 의미한다. 각 가중치는 실험을 통하여 결정된다. a_1, a_2, a_3 는 각각의 요소의 가중치를 의미한다.

<주석유사도 알고리즘>
-Input :
두 개의 문서 혹은 데이터에 대한 주석문서 A_a ,A_b
E_i는 (A_a의 요소), E_j는 (A_b의 요소), i=1..n, j=1..m,
가중치 값 a1, a2, a3
-Output : 주석문서의 유사도 $S(A_a,A_b)$ 의 값

ExactMatchCount = 0
PartiallyMatchCount = 0
ValueMatchCount = 0
for each item in A_a i=1..n
  for each element in A_b j=1..m

```
            if Exactmatch(Ei,Ej) then
            ExactMatchCount = ExactMatchCount +1
            else if Partiallymatch(Ei,Ej) then
          PartiallyMatchCount =
              PartiallyMatchCount +1
            else if Valuematch(Ei,Ej) than
            ValueMatchCount = ValueMatchCount +1
    end for
end for
S(Aa,Ab) = a1 · ExactMatchCount
          + a2 · PartiallyMatchCount
          + a3 · ValueMatchCount
return
```

4. 실험 평가

본 논문에서 제안한 시맨틱 주석 문서의 유사도 함수를 평가하기 위해 <표 3>에서 제시한 3개의 영화와 관련된 9개의 주석 문서를 대상으로 실험을 진행하였다. 영화1과 영화2는 같은 감독이 연출했다는 공통점을 가지고 있고 영화2와 영화3은 주연 배우가 같다는 공통점을 가지고 있다.

<표 3> 실험 대상 영화 분류

분류 \ 특성	제목	감독
영화1	아바타	제임스 캐머런
영화2	타이타닉	제임스 캐머런
영화3	캐치 미 이프 유 캔	스티븐 스필버그

각 영화별로 <표 4>와 같은 특성을 가지는 3개의 주석 문서를 RDF/XML로 표현하여 실험에 사용하였다. 모든 주석 문서는 더블린코어, FOAF, MPEG-7 스키마와 직접 설계한 영화 스키마를 사용하여 작성하였다.

<표 4> 실험 대상 주석 문서의 특성

분류 \ 특성	대상	내용
주석1	텍스트 문서	영화의 줄거리 또는 비평
주석2	이미지	영화의 포스터
주석3	동영상	영화의 예고편

본 논문에서 제안한 시맨틱 주석 문서의 유사도 측정 알고리즘은 C 언어로 구현하였고 1GB RAM, 윈도우 XP 운영체제가 설치된 3.4GHZ Pentium 4 PC 환경에서 실험하였다.

유사도 측정 함수의 3개 가중치 값을 결정하기 위해 <표 5>와 같이 각 가중치를 변화시켜 실험을 진행하였다.

<표 5> 실험별 가중치 값

분류 \ 가중치	a_1	a_2	a_3
실험1	0.2	0.3	0.5
실험2	0.3	0.2	0.5
실험3	0.2	0.5	0.3
실험4	0.3	0.5	0.2
실험5	0.5	0.2	0.3
실험6	0.5	0.3	0.2

<표 3>에서 제시한 3개의 영화와 관련하여 작성된 9개의 시맨틱 주석 문서에 대해 <표 5>에서 제시한 가중치로 실험을 진행한 결과 모든 실험에서 주석 문서의 대상과 관계없이 같은 영화에 대한 주석 문서들 간의 유사도가 높게 측정되었다. 특히, 다른 실험들에 비해 실험6에서 유사도가 높게 측정되었다.

<표 6>은 <표 5>의 6개의 실험중에서 가장 높은 유사도를 보인 실험6의 가중치를 가지고 영화1의 주석 문서1과 다른 주석 문서들 간의 유사도를 측정한 결과를 보여준다. <표 6>에서 주석 문서의 대상과 관계없이 같은 영화에 대한 주석 문서들 간의 유사도가 상대적으로 높게 평가된 것을 확인할 수 있다. 영화1의 줄거리와 비평 내용을 담고 있는 텍스트 문서에 대한 주석 문서1은 8.5의 유사도가 측정된 같은 영화의 예고편 동영상에 대한 주석 문서2와 가장 유사하다고 판단할 수 있다. 그리고 두 번째로 높은 6.4의 유사도가 측정된 주석 문서3도 같은 영화의 포스터 이미지에 대한 주석 문서이다. <표 6>에서 영화1의 주석 문서1은 영화3의 주석 문서들에 비해 영화2의 주석 문서들과 비교적 높은 유사도를 가지고 있음을 확인할 수 있다. 이것은 영화1과 영화2가 같은 감독이 연출한 공통점을 가지고 있어서 일치하는 값을 가지는 공통적인 요소와 속성이 영화3에 비해 많이 존재하기 때문이다.

<표 6> 영화1-주석1과 다른 주석 문서들 간의 유사도 측정 결과

영화1		영화2			영화3		
주석2	주석3	주석1	주석2	주석3	주석1	주석2	주석3
6.4	8.5	4.3	1.5	3.0	3.0	1.0	1.8

<표 7>은 실험6에서 영화2의 주석 문서2와 다른 주석 문서들 간의 유사도를 측정한 결과를 보여준다. <표 6>과 마찬가지로 <표 7>에서 주석 문서의 대상과 관계없이 같은 영화에 대한 주석 문서들 간의 유사도가 상대적으로 높게 평가된 것을 확인할 수 있다. 영화2의 포스터 이미지에 대한 주석 문서2는 5.3의 유사도가 측정된 같은 영화의 예고편 동영상에 대한 주석 문서3과 가장 유사하다고 판단할 수 있다. 그리고 두 번째로 높은 4.0의 유사도가 측정된 주석 문서2도 같은 영화의 줄거리와 비평을 담고 있는 텍스트 문서에 대한 주석 문서이다.

<표 7> 영화2-주석2와 다른 주석 문서들 간의 유사도 측정 결과

영화1			영화2		영화3		
주석1	주석2	주석3	주석1	주석3	주석1	주석2	주석3
2.6	3.9	3.9	4.0	5.3	3.1	3.9	3.1

<표 8>은 실험6에서 영화3의 주석 문서3과 다른 주석 문서들 간의 유사도를 측정한 결과를 보여준다. <표 6>이나 <표 7>과 마찬가지로 <표 8>에서도 주석 문서의 대상과 관계없이 같은 영화에 대한 주석 문서들 간의 유사도가 상대적으로 높게 평가된 것을 확인할 수 있다. 영화3의 예고편 동영상에 대한 주석 문서3은 6.1의 유사도가 측정된 같은 영화의 텍스트 문서에 대한 주석 문서1과 가장 유사하다. 그리고 두 번째로 높은 4.6의 유사도가 측정된 주석 문서2도 같은 영화의 포스터 이미지에 대한 주석 문서이다. <표 8>에서 영화3의 주석 문서3은 영화1의 주석 문서들에 비해 영화2의 주석 문서들과 비교적 높은 유사도를 가지고 있음을 확인할 수 있다. 이것은 영화2와 영화3이 같은 배우가 주연한 공통점을 가지고 있어서 일치하는 값을 가지는 공통적인 요소와 속성이 영화1에 비해 많이 존재하기 때문이다.

<표 8> 영화3-주석3과 다른 주석 문서들 간의 유사도 측정 결과

영화1			영화2			영화3	
주석1	주석2	주석3	주석1	주석2	주석3	주석1	주석2
1.5	1.8	2.6	3.8	3.6	4.3	6.1	4.6

<표 6>, <표 7>, <표 8>에서 이미지에 대한 주석 문서에 비해 텍스트 문서에 대한 주석 문서와 동영상에 대한 주석 문서 사이에 더 높은 유사도가 측정되었다. 이것은 영화 전체에 대한 줄거리와 비평을 담고 있는 텍스트 문서와 영화 전체의 예고편 동영상에 대한 주석 문서는 영화의 한 장면과 같이 정지된 하나의 이미지를 대상으로 하는 주석 문서에 비해 같은 값을 가지는 공통의 요소와 속성을 많이 가지고 있기 때문이다.

실험을 통해서 알 수 있듯이, 제안된 유사도 측정 알고리즘을 통하여, 웹상의 텍스트, 이미지, 동영상 등의 자료들을 주석 문서로 지니고 있는 경우에 자료의 형태와 상관없이 유사한 주제를 지니고 있는 자료를 검색하는데 사용될 수 있음을 알 수 있다.

5. 결론

인터넷상의 수많은 자료들을 좀 더 정확하고 간결하게 검색하고자 하는 욕구가 많아질수록 기존의 HTML 문서에 대한 키워드 검색의 범주를 벗어나, XML 기반의 메타데이터정보구축을 통한 차세대 검색 시스템을 개발하고자 하는 연구가 활발히 진행되고 있다. 본 연구에서는 시맨틱 주석으로 구성된 문서들 간의 유사도 측정을 통하여 좀 더 관계가 많은 문서들을 찾고 이를 바탕으로 검색 결과를 산출하도록 하는데 연구의 중점을 두었다. 다양한 주석 문서를 내용 분석에 따라 단계별 형태로 분류하고, 상이한 종류의 자료 간에도 검색이 가능하도록 주석 문서간의 유사도를 측정하는 방법을 제시하였다. 유사도 관계를 측정하기 위한 항목으로 스키마, 요소, 속성, 값, 단계의 5가지 항목을 사용하였다.

실험을 통하여 논문에서 제안한 유사도 관계 측정 알고리즘을 결과 값들을 통하여 유사한 주제의 주석 문서들 간의 유사도가 높음을 보였다. 제안된 알고리즘은 차세대 인터넷 검색 시스템에 적용 가능할 것으로 보이며, 추후 연구로는 좀 더 다양한 형태의 많은 자료와 주석에 대한 추가적인 연구를 통하여 대용량 데이터 환경에서의 적합성을 보일 것이다.

참 고 문 헌

[1] Siegfried Handschuh and Steffen Staab, editors. Annotation for the Semantic Web. IOS Press, 2003.

[2] Lawrence Reeve and Hyoil Han, The Survey of Semantic Annotation Platforms, The 20th Annual ACM Symposium on Applied Computing (ACM SAC) 2005, Santa Fe, New Mexico, 2005

[3] "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup", Maria Vargas-Vera, Enrico Motta, John Domingue, Mattia Lanzoni, Arthur Stutt and Fabio Ciravegna, The 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Springer Verlag, 2002

[4] PLUS Group, "Simple HTML Ontology Extenstions" http://www.cs.umd.edu/projects/plus/SHOE, 2002.

[5] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M., KIM - Semantic Annotation Platform in 2nd International Semantic Web Conference (ISWC2003), 2003, pp 834-849.

[6] The Dublin Core Metadata Initiative, Dublin Core Metadata Element Set, Ver. 1.1: Reference Description. http://dublincore.org/documents/2010/10/11/dces/

[7] Visual Resources Association Data Standards Committee, VRA Core Categories, Version 3.0. http://www.vraweb.org/about/index.html

[8] Information Technology - Multimedia Content Description Interface (MPEG-7). Standard No. ISO/IEC 15938:2001, International Organization for standardization(ISO), 2001.

[9] SF. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. IEEE Trans. on Circuits and Systems for Video Technology, 11(6) , June 2001, pp 688-695

[10] Stephan Bloehdorn, Kosmas Petridis, Carsten Saathoff, Nikos Simou, Vassilis Tzouvaras, Yannis Avrithis, Siegfried Handschuh, Yiannis Kompatsiaris, Steffen Staab, Michael G. Strintzis, "Semantic Annotation of Images and Videos for Multimedia Analysis", The Semantic Web: Research and Applications: Proceedings of the Second European Semantic Web Conference, ESWC 2005, pp 592-607

[11] Image Annotation on the Semantic Web, W3C Working Draft 22 March 2006
http://www.w3.org/TR/2006/WD-swbp-image-annotation-20060322/

김병곤

1990년 : 홍익대학교 공과대학 전자계산학과 이학사
1992년 : 홍익대학교 공과대학 전자계산학과 이학석사
2001년 : 홍익대학교 공과대학 전자계산학과 이학박사
1992년~1998년 : 국방과학연구소 연구원
2001년~현재 : 부천대학 e-비즈니스과 부교수
관심분야 : DB, 데이터웨어하우스, 시맨틱웹

오성균

1981년 : 홍익대학교 이공대학 전자계산학과 이학사
1984년 : 연세대학교 산업대학원 전자계산학과 공학석사
1999년 : 홍익대학교 공과대학 전자계산학과 이학박사
1987년~현재 : 서일대학 소프트웨어과 교수
관심분야 : 능동데이터베이스, XML모델링, 소프트웨어공학