질의어 자동수정을 이용한 메타시소러스 검색 방법

김종광[O]* 하원식* 김태용* 류중경* 이정현**
인하대학교 전자계산공학과*, 인하대학교 컴퓨터공학부**
{be3light[O], sigboy, tykim jkryu}@nlsun.inha.ac.kr*, jhlee@inha.ac.kr**

The Method of Searching Metathesaurus, Using Automatic Modified a Query

Jonggwang Kim[O]* Wonsik Ha* Taeyong Kim* Joongkyung Ryu* Junghyun Lee**
{Dept.*, School**} of Computer Science & Engineering, INHA University

요 약

UMLS(2003AA edition 기준)의 메타시소러스는 다국어를 지원하며 875,233개의 개념(concept)과 2,146,897개의 개념명(concept name)을 포함한다. 현재 UMLS 메타시소러스 검색을 제공하는 PubMed나 NLM에서는 UMLS에서는 개념명에 존재하지 않는 잘못된 질의나, 잘못된 구문 또는 개념명의 일부를 이용한 검색이 불가능하다. 이는 사용자가 UMLS에서 정보를 얻기 위해서는 정확한 의학용어를 숙지해야 되며, UMLS 메타시소러스의 데이터가 잘못 되었을 경우 정보를 얻을 수 없다. 본 연구에서는 이러한 문제점을 보완하기 위해서 자연어처리에서 연구되고 있는 문자열 간의 유사도 측정방식을 적용하여 잘못된 질의어에 대한 자동수정 기능을 이용한 메타시소러스 검색방법을 제안한다. 제안한 방법에서는 질의어를 자동수정하기 위하여 철자사전을 자동으로 추출하고 문자열 비교알고리즘을 도입하여 질의어와 철자사전간의 용어의 유사도를 측정한다. 유사도에 의하여 얻어진 용어를 메타시소러스의 형식에 맞게 변환하여 질의에 대한 최적의 결과를 얻을 수 있도록 한다. 제안된 방법의 성능을 평가하기 위해서 최근(2003년 8월) bi-gram 방식을 도입한 NLM에서의 시스템과 비교 평가한다.

1. 서 론

UMLS(2003AA edition 기준)의 의학용어는 다국어를 지원하며, 2,146,897개의 개념명을 포함한다. UMLS 검색의 대표적인 시스템인 기존의 미국 국립의학 도서관(NLM) 검색시스템에서의 UMLS 메타시소러스 검색에 문서에 없는 잘못된 질의나 잘못된 구문 또는 개념명의 일부를 이용한 검색이 어렵다. 예를 들어 "oculus"를 찾을 때 "oculu" 또는 "aculus" 등의 잘못된 입력일 경우는 "**Query unsuccessful.**"이라는 출력과 함께 검색을 해내지 못한다. 이럴 경우 의학용어에 익숙지 않은 사용자나 다른 언어로 된(Dutch, French, Finnish, German, Italan, Portuguese, Russian, Spanish 등) 정보를 검색하기 위하여 정확한 전문용어를 알고 있어야 한다. 또는 UMLS 메타시소러스를 개발하는 개발자의 실수로 인하여 잘못된 정보를 입력 시에는 그 정보를 검색하지 못하는 결과를 초래하게 된다. 2003년 8월부터 NLM에서는 문자에 기반한 bi-gram 접근 방법 등을 통하여 전문용어에 대해 철자 오류를 검출하였다[1]. 본 연구에는 UMLS 메타시소러스 검색시스템에 잘못된 질의어를 자동수정하는 기능을 가지는 새로운 방법을 제안하며 NLM의 방식과 성능을 비교 평가한다.

2. 관련연구

UMLS(Unified Medical Language System)는 동일한 개념에 대한 용어표현차이로 인한 정보의 검색 및 통합문제를 해결하기 위하여 NLM에서 개발된 통합의학언어시스템이다. UMLS는 개념을 다루는 메타시소러스(Metathesaurus)와 모든 개념에 대한 그룹화 및 개념간 관계를 구축해 놓은 의미망(Semantic Network) 자연어 처리를 위해 개발된 전문가사전(Specialist Lexicon) 등을 포함한다.

PubMed(http://ncbi.nlm.nih.gov/Pub/Med/) 와 NLM Gateway (http://gateway.nlm.nih.gov/gw/Cmd)에서 웹을 통하여 메타시소러스의 검색을 제공하며, 검색시스템에 최대한 이용자의 요구를 수용하여 보다 적극적인 이용자 중심의 검색시스템을 개발하려고 노력하고 있다[2].

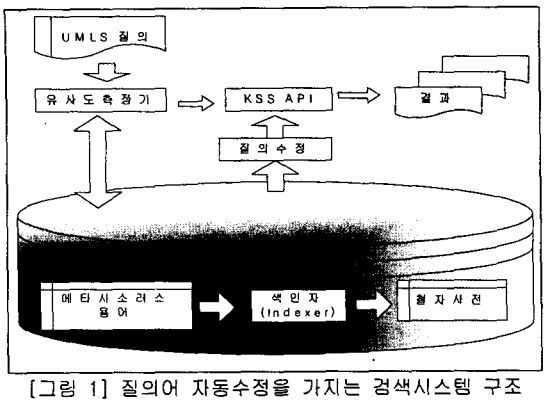
메타시소러스는 "동일한 개념의 서로 다른 명칭과 관점을 함께 연결하고 상이한 개념 사이의 유용한 관계를 밝히는 것"을 목적으로 하고 있다. 이런 목적을 성취하기 위하여 메타시소러스는 개념, 용어, 문자열의 3단계 체제를 가지고 있다. 실제 자료에 나타나는 형태는 문자열이라고 부르고 단순히 철자적 변형에 불과한 문자열끼리는 같은 용어로, 의미하는 바가 같은 용어는 같은 개념으로 규정하였다. 이와 같이 규정된 개념, 용어, 문자열을 식별하고 연결해주기 위해서 각각 고유개념식별기호(CUI), 공통용어식별기호(LUI), 고유문자열식별기호(SUI)를 부여하고 개념과 용어 사이의 다대다 대응 관계를 식별기호 사이의 연결구조로 표현하였다. 온톨로지는 개념화 된 것을 명확화한 것으로 도메인의 어휘로 정의되고 어휘내의 용어(term)의 사용으로 제약한다[2]. 문자열은 문자 숫자식의 순서를 가지며 각각은 공백 또는 구두점 등으로 분리된다. 사용자들이 UMLS에서 정보를 검색할 때 고유개념식별기호(CUI) 또는 용어를 이용한다.

UMLS에서 검색방법은 주로 CUI 또는 용어를 이용한다. 하지만 실제로 사용자는 UMLS의 CUI가 아닌 용어를 이용한 검색을 한다. 용어를 이용하여 CUI를 찾고 그에 따른 LUI와 SUI를 검색함으로써 정보를 얻을 수 있다. 용어를 이용한 CUI는 메타시소러스 중에서 MRCON테이블과 MRSO테이블을 이용하여 검색가능하다. 그러나 MRCON 테이블은 전체 2,146,897건의 레코드를 가지고 있으므로 일반적 검색방법으로는 검색시 많은 시간이 소요되므로 검색을 위한 색인화 및 재배치 등의 방법과 빠른 검색알고리즘을 이용한 여러 가지 방법들이 시도되고 있다[3][4].

3. 질의어 자동수정 기능을 이용한 UMLS에서의 의학용어 검색

3.1 검색시스템 구성

그림1 은 본 연구에서 제안한 메타시소러스에서 유사도를 이용한 검색시스템의 구조를 나타낸다.



[그림 1] 질의어 자동수정을 가지는 검색시스템 구조

본 연구에서 제안한 검색시스템은 유사도 측정기와 철자사전을 이용한다. 색인자는 메타시소러스의 용어에서 불용어와 중복용어 등을 제거하는 기능을 하며, 용어의 개수에 따른 가중치를 부여하여 철자사전을 만든다. 입력된 질의어에 대한 일치하는 용어가 존재하면 UMLS에서 제공되는 KSS(Knowledge Source Server) API를 이용하여 질의어에 대한 결과를 출력하고 일치하는 결과가 없으면 유사도 측정기와 철자사전을 이용하여 질의어를 수정한 후 질의에 대한 결과를 찾는다.

3.2 잘못된 검색데이타의 수정을 위한 철자사전 구축

본 연구를 위해 NLM에서 라이센스를 취득하였고 UMLS의 네가지 구성요소 중 한 컴포넌트인 2003AA edition을 이용하였으며, Windows2000과 RedHat Linux9.0의 운영체제에 각각 VisualBasic 6.0과 Java를 이용하여 검색엔진을 구축하였다. UMLS에는 자연어 처리를 위하여 약 2만 단어와 1만2천여개의 일반어휘에 관한 정보를 포함하는 전문가사전을 제공하지만 본 연구에서는 의학용어 검색에 대한 질의어수정 기능의 성능을 높이기 위하여 전문가사전을 이용하지 않고 MRCON으로부터 직접 MRCHK라는 철자사전을 알고리즘 1의 철자구축 알고리즘을 이용하여 추출하였다. MRCHK는 CUI 순으로 정렬된 MRCON에서 2,146,897개의 개념명을 추출하고, 불용어와 중복되는 용어를 제거하여 357,266용어를 포함하는 철자사전과, 각 용어에 대하여 계산된 가중치를 포함한다. 철자사전의 추출결과는 4장의 실험 및 성능평가의 그림 3에서 보여준다.

```
Open "MRCON" For Input As #1
 Do While Not EOF(FileNum)
  Line Input #1, strInput
          IStr = Split(strInput, "|") '개념분리
           Dstr = RemoveStopWord(IStr(6))
             '불용언, 구문기호 제거/ 분리
           For J = 0 To UBound(Dstr) - 1
               Ostr=ComDic(Dstr(J), ByVal Count)
                'MRCHK에 있는 용어와 비교/가중치 계산
               Pos=FindPos(Ostr) '용어의 삽입위치를 결정
               Call OutPut("MRCHK", Pos, Ostr, Count)
                '철자사전에 용어와 가중치 입력
           Next J
           DoEvents
        rc_cnt = rc_cnt + 1
 Loop
```

[알고리즘 1] 철자구축 알고리즘

3.3 잘못된 질의어의 자동수정 및 동의어 처리

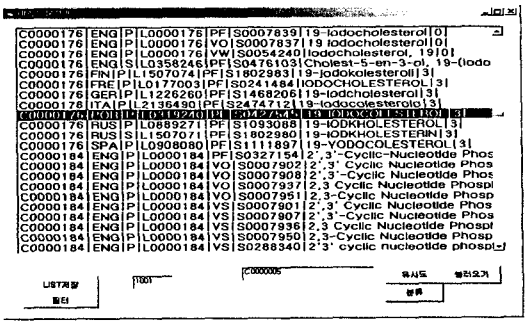
메타시소러스의 각각의 용어는 고유한 문자열식별자인 SUI를 가지며 이는 실제 자료에 나타나는 형태로서 철자변형, 단복수변형 및 언어변형에 따라서 고유한 SUI가 부여된다.

입력 받은 질의어에서 X번째 질의가 $O_{x1}, O_{x2}, \ldots, O_{xn}$ $(O_\xi, 1 \leq i \leq n)$ 이고 Y번째 철자사전 데이타가 $O_{y1}, O_{y2}, \ldots, O_{ym}(O_\xi, 1 \leq j \leq n)$ 일때, 각각의 최대유사도를 $Q_{x1}, Q_{x2}, \ldots, Q_{xn}, Q_{y1}, Q_{y2}, \ldots, Q_{ym}$라고 하면 X, Y의 유사도는 식(1)에 의해 계산되어 진다.

$$SIM(X, Y) = \frac{\sum_{i=1}^{n} Q_{xi} + \sum_{j=1}^{m} Q_{yj}}{2(n+m)} \qquad \text{식(1)}$$

본 연구에서 질의어 수정에는 다음을 고려하였다.

1. 질의는 단일명사인 경우도 있지만 복수개의 단어를 가지는 문자열일 수도 있다. 질의를 각각의 단어로 분리하고 각각의 단어들을 철자사전에서 유사도가 가장 높은 단어로 대체한다.
2. 메타시소러스에서는 철자적 변형에 의한 다른 문자열을 같은 용어로 규정한다. [그림 2]는 MRCON 데이타의 일부를 보여준다. MRCON은 각각 CUI(개념식별자)|LAT(용어의 언어)|TS(용어의 상태)|LUI(공통용어 식별자)|STT(문자열 유형)|SUI(문자열 식별자)|STR(문자열)|LRL(최소 제한 수준)| 이다. 그림2를 보면 19-iodocholesterol과 19 iodocholesterol, Iodocholesterol, 19가 같은 LUI(L0000176)를 가지므로 이를 공통용어로 인식한다. 이는 메타시소러스 검색에는 특수문자나 숫자 등을 고려할 필요가 없음을 보여준다.



[그림 2] MRCON의 구조

3.4 질의어의 자동수정 기능을 가지는 의학용어 검색시스템 알고리즘

본 검색시스템은 알고리즘2의 질의어 자동수정을 통한 검색을 이용한다.

1. 입력 : n개의 문자열로 구성된 질의어 Φ = {t1, t2, ..., tn}
2. KSS API를 이용하여 질의어를 검색한다.
3. 질의어 검색이 성공이면 5로 간다.
   질의어 검색이 실패이면 ti ~ tn까지 철자사전의 데이터와 유사도를 구한다. (유사도가 높은 순으로 스택에 저장한다. 이 때 유사도가 같으면 가중치가 높은 용어를 우선적으로 저장한다.)
4. ti ~ tn까지 유사도가 1이 아니면 스택의 가장 상위의 용어로 대체하고 2로 간다. 스택에서 대체된 용어를 삭제한다.
5. 검색결과에 따른 CUI, 동의어 및 MRDEF에 정의된 CUI에 대한 설명을 출력한다.

[알고리즘 2] 질의어 자동수정을 통한 검색

4. 실험 및 성능평가

MRCON으로부터 5,418,480개의 어절을 분리했다. 불용언 및 중복사전을 제거하여 357,266의 철자사전을 추출했다. 그림 3은 추출된 철자사전 데이타의 일부이다. 용어 옆의 숫자는 용어의 중복수이고 '*' 옆의 숫자는 중복된 용어의 수로 가중치를 나타낸다.

.	.
.	.
.	.
Atrophica*1	parryana*1
Atrophicus*1	parryi*13
Atrophie*9	parryii*2
Atrophien*2	pars*135
Atrophies*52	.
.	.
.	.
.	straal*1
DENTALES*40	strab*1
DENTALIS*1	strabismic*2
DENTARIA*138	strabismus*57
DENTARIAS*17	stracheyi*1
DENTARIO*47	.
.	.
.	.
.	zytotoxische*1
	zytotoxisches*1
	weight*7

[그림 3] 철자사전 데이타

질의어 "cald"일 경우에는 그림4에서 보는 것과 같이 "SCALD"와 "scald"는 같은 동일한 유사도를 가진다. "scald"는 "SCALD"보다 높은 가중치를 가지므로 확률적으로 "scald"가 우선한다. 본 연구에서는 "cald"의 질의어를 "scald"로 대체하고 질의를 수행한다.

ALD	0.857142857142857
CAL	0.857142857142857
CALLED	0.8
Cad	0.857142857142857
Cal	0.857142857142857
[redacted]	[redacted]
ald	0.857142857142857
cal	0.857142857142857
caldo,	0.8
caldus	0.8
calida	0.8
cld	0.857142857142857
[redacted]	[redacted]
scald,	0.8
scaled	0.8

[그림 4] "cald"의 질의 결과

그림5를 보면 질의어 "cold"에 대하여 유사도 "1"을 가지는 결과가 존재함을 볼 수 있다. 이런 경우에는 가중치에 관계없이 질의어가 잘못되지 않았으므로 질의어를 수정할 필요가 없다.

(COLD)	0.8
COD	0.857142857142857
COL	0.857142857142857
COLADA	0.8
COLADO	0.8
[redacted]	[redacted]
COLD,	0.888888888888889
COOLED	0.8

Cod	0.857142857142857
Coiled	0.8
Col	0.857142857142857
[redacted]	[redacted]
Cold,	0.888888888888889
Colds	0.888888888888889
OLD	0.857142857142857
Old	0.857142857142857
cld	0.857142857142857
cod	0.857142857142857
coiled	0.8
col	0.857142857142857
[redacted]	[redacted]
cold)	0.888888888888889
cold,	0.888888888888889
cold;	0.888888888888889
cold]	0.888888888888889
could	0.888888888888889

[그림 5] "cold"의 질의 결과

본 연구의 출발은 NLM에서 메타시소러스 검색에 질의어 자동수정 기능을 가지지 않는 것에 착안하여 연구를 시작하였으나 2003년 8월부터 NLM에 질의어 자동수정기능이 보완 되었다. 표1과 같이 본 연구와 NLM의 질의어 자동수정기능을 비교하여 보았을 때 본 연구가 좀더 원하는 결과를 잘 찾아냄을 볼 수 있다.

[표 1] NLM Gateway와의 비교

질의어	NLM Gateway	질의어 수정을 통한 검색
cald	caldo	scald
cold	cold	cold
omeodomain	oleuropein opiocortin	HOMEODOMAIN homeodomain
aculus	Acavus	jaculus
abdome egudo	ABDOME AGUDO	ABDOME NEGUNDO
clod	검색실패	Cloud rat
Sllep Apea	검색실패	Sleep Apnea

5. 결론

본 연구에서는 의학용어 검색시 잘못된 질의에 대하여 메타시소러스에의 의학용어를 추출하고 철자사전을 구축한 후 잘못된 질의어에 대한 자동수정을 함으로써 사용자가 보다 편하게 의학정보를 접근하게 하였다. 본 연구에서는 의학용어에서 추출한 자체 사전을 이용함으로써 신뢰도를 높혔고 사전을 최소화함으로써 검색 속도가 향상되었다.

향후 과제로는 보다 많은 철자검증 알고리즘을 도입해서 사용자가 원하는 질의어를 빠른 접근과, 한글 또는 일본어 중국어 등의 의학용어를 메타시소러스에 추가하여 이에 대한 본 연구의 적용으로 더욱 성능을 향상시킬 수 있을 것이다.

6. 참고문헌

[1] Guy Divita, Allen C. Browne, Tony Tse, et. al., "A Spelling Suggestion Technique for Terminology Servers," National Library of Medicine, 2003.

[2] Suarez HH, Hao X, and Chang IF, "Searching for information on the Internet using the UMLS and Medical World Search," In *Proceedings of the 1997 Annual AMIA Fall Symposium*. Nashville, TN: Hanley & Belfus pp. 824-828, 1997.

[3] UMLS Knowledge Sources. (14th ed.) Bethesda(MD): National Library of Medicine 2003AA, pp. 1-102, 2003.

[4] X. Qi, S. Sung, Z. Li, C, Lu and P. Sun, "Faster Algorithm of String Comparison," Journal of Pattern Analysis and Applications (accepted for publication), 21 Dec 2001.