http://dx.doi.org/10.7236/JIIBC.2013.13.1.83
JIIBC 2013-1-12

흰개미 군집 알고리즘을 이용한 유사 블로그 추천 시스템에 관한 연구

A Study of Similar Blog Recommendation System Using Termite Colony Algorithm

정기성, 조이석, 이말례*

Gi Sung Jeong, I-seok Jo, Malrey Lee

요 약 본 연구의 목적은 유사 블로그 추천 시스템을 통해서 특정 주제의 유사도에 따라 주제를 찾아 주는 것이다. 유사 추천 시스템을 실현하기 위해서는 대규모 데이터 집합에서 유사항목을 가진 그룹을 찾을 수 있도록 군집해야 한다. 군집화(clustering) 기법은 군집하고자 하는 목적에 따라 적합한 기법과 군집수가 결정되어야 한다. 군집기법으로는 가장 많이 사용되는 K-means 알고리즘을 사용 하였고 추천 알고리즘은 흰개미 군집 알고리즘을 사용하였다. 흰개미 습성 모델을 이용한 군집화 기법은 K-means 알고리즘이 갖고 있는 적절한 군집 갯수 문제점을 해결하고, 군집화 시간을 단축하며, 군집을 위한 군집 평균 이동횟수를 개선한다.

Abstract This paper proposes a recommending system of the similar blogs gathered with similarities between blogs according to the similarity, dividing words, for each frequency, that individual blogs have. It improved the algorithm of k-means, using the model of the habits of white ants for better performance of clustering, and showed better performance of clustering as a result of evaluating and comparing with the existing algorithm of k-means as the improved algorithm. The recommending system of similar blog was designed and embodied, using the improved algorithm. TCA can reduce clustering time and the number of moving time for clustering compare with K-means algorithm.

Key Words : blog, termite, k-means, clustering, recommendation system

Ⅰ. 서 론

오늘날 블로그를 통해 정보를 공유할 수 있지만, 각각 특성이 다른 블로그에서 단순 키워드 검색만을 통해서는 블로그의 특성을 찾을 수 없다. 따라서 본 논문에서는 수없이 산재된 블로그 중에서 본인이 찾고자 하는 특정 주제에 맞는 블로그를 찾기 위해서, 각 블로그의 단어를 빈도별로 나누어 유사도를 측정하고 이를 기준으로 유사 블로그를 군집(cluster)하여, 블로그 특성 군집을 찾을 수 있는 방안을 제안 하였다. 또한 군집 그룹 속에서 특정 주제의 유사도에 따른 순위를 정하는 유사 블로그 추천 시스템을 제안하였다.

*정회원, 전북대학교 컴퓨터공학과, 교신저자
접수일자 2013년 1월 3일, 수정완료 2013년 2월 5일
게재확정일자 2013년 2월 8일

Received: 3 January 2013 / Revised: 5 February 2013 / Accepted: 8 February 2013
*Corresponding Author: mrlee@chonbuk.ac.kr
Dept. Computer Science & Engineering, Chonbuk National University, Korea

유사 블로그 추천 시스템을 통해서 특정 주제의 유사도에 따라 랭크 하는 것이 본 논문의 목적이다. 따라서, 본 논문의 목적을 달성하기 위해서는 대규모 데이터 집합에서 유사항목을 가진 그룹을 찾을 수 있도록 군집해야 한다. 군집화(clustering) 기법은 군집하고자 하는 목적에 따라 적합한 기법과 군집수가 결정되어야 한다[1,2].

본 논문의 구성은 다음과 같다. 먼저 2장에서는 군집화 알고리즘에 대해서 설명하였다. 3장에서는 흰개미의 습성을 소개하고 그 특성을 이용한 군집화 기법을 소개한다. 4장에서는 전체적인 시스템 구성과 설계도를 작성하여 실제 단어에 의한 추천 시스템의 실행된 구현 결과와 함께 제안한 흰개미 습성 모델 알고리즘과의 실험평가를 한다. 5장에서 결론과 향후 연구 과제를 제시한다.

II. 군집화 알고리즘

군집화 기법 중에 많이 사용되는 기법은 k-means 알고리즘으로 분할군집화 기법의 일종이다.

1. k-means 알고리즘

직관적인 이해가 쉽고, 구현이 간편해서 많이 사용되는 기법인데, 초기 군집 개수를 설정해야 한다. k-means 알고리즘은 [표 1]과 같다. 임의의 초기 군집 위치에서 각 샘플들의 거리를 측정한 후 가까운 곳에 샘플의 군집을 형성하는 방식이다. 즉, 군집된 샘플들의 중심으로 군집점을 이동시킨 후 군집 중심이 바뀌지 않을 때까지 반복하는 알고리즘이다[1,2].

표 1. k-means 알고리즘
Table 1. K-means Algorithm

```
입력: 샘플 집합 X={x1, x2, ..., xN}, 군집의
      개수 k
출력: 군집 해 C
begin:
  k개의 군집 Z={z1, z2, ..., zN}를 초기화
  while (TRUE) {
    for (i=1 to N)
     // xi를 가장 가까운 군집 중심에 배치
    if (이 배정이 이전 루프의 배정과 같음) break;
    for (j=1 to k)
    // zj에 배정된 샘플의 평균으로 zj를 대치하며,
  } //  군집점 이동
end:
```

III. 흰개미 군집 알고리즘

생물학 분야에서는 무리지어 생활하는 생물들의 연구가 활발하게 이뤄지고 있다. 즉, 개미, 벌, 새, 물고기 같은 무리 생활을 하는 동물들을 관찰하고 그들의 습성을 찾는 것이다. 이러한 생물들의 습성을 이용한 새로운 알고리즘이 만들어져 활용되고 있다. 특히, 개미를 이용한 알고리즘(Ant Conoly Algorithm;ACA)과 무리떼의 습성을 이용한 알고리즘(Particle Swarm Algorithm;PSA) 등의 많은 연구가 되고 있다. 본 논문에서는 군집 습성 알고리즘에 해당하는 흰개미 군집의 먹이 탐색 습성을 이용한 흰개미 군집 알고리즘(Termite Colony Algorithm ;TCA)을 제안한다[3,4,5,6,7].흰개미 군집 알고리즘은 흰개미 행동습성의 확률값을 기반으로 설계한다. 흰개미 군집 알고리즘을 통해서 k-means 군집에 필요한 적절한 초기 군집해를 찾는 것이 목적이다.흰개미 탐색을 통해서 군집 속 샘플 x_{ij}의 밀도를 추정할 수 있으며, 밀도에 맞게 추정된 초기해는 군집화 성능에 영향을 준다.

1. TCA 탐색 절차를 위한 가정

가정 1. 흰개미 초기 위치는 시뮬레이션 절차상 2차원 좌표의 중앙에 위치시킨다.

가정 2. 흰개미 무리를 6개 집단으로 각각 53±9º 각을 배정한다.

가정 3. 각각의 무리들의 방향에서 [그림 1]과 같은 확률이 갖도록 랜덤값을 부여한다

가정 4. 검색 범위의 빠른 처리를 위해서 무리를 확대시킨다.

가정 5. 확대범위에 따라 10번 이내로 수행을 마친다.

가정 6. 중심점과, 최고의 빈도를 갖는 무리의 좌표를 초기 좌표로 지정한다.

가정 7. 기본적으로 흰개미 무리수는 6개가 된다. 가정 2의 배정된 각이 작을 경우 1개의 무리가 추가될 수 있다.

그림 1. 흰개미 군집의 방향 선택 확률
Fig. 1. Probability of direction slelection in Termite Colony

2. 흰개미 군집 알고리즘

흰개미 군집은 가정된 임의의 배정각에 의해서 위치가 결정된다. 흰개미 군집이 경로를 탐색하는 과정에서 특정 무리가 샘플을 $x_{|(X_n/6)/2|}$ 개를 만나게 되면 탐색을 마치고 그 지역의 중심점을 초기해로 배정한다. 이로 인해서 밀도가 높은 지점에 초기해를 갖게 된다.

표 2. 흰개미 군집 알고리즘
Table 2. Termite colony Algorithm

```
입력: 임의의 흰개미 무리값 설정
출력: 초기 군집 해 Z={z1, z2, ..., z6} // 군집 초기 위치 값
begin :
 x_mean = max(x_i) − min(x_i)/2
// 전체 샘플 범위의 중심점을 찾는다.
 y_mean = max(y_i) − min(y_i)/2
 while (TRUE) {
   Termite_Search_Path()
// [그림 3-4]의 확률 경로로 탐색
   if ( t_i=x_|(X_n/6)/2| ) break;
//  t_i : 흰개미 군집
 } //  x_|(X_n/6)/2| : 샘플 밀도

 for (j=1 to 6) {
   Swap(z_j,t_j)
// z_j에 흰개미 종료 지점 t_j를 대치.
 }
 z_7 = x_mean,y_mean
// z_7에 전체 샘플 범위의 중심점을 할당.
 Initial_Z()
// k개의 군집 Z={z1, z2, ..., z7}를 초기화
 while (TRUE) {
   for (i=1 to N)
// x_i를 가장 가까운 군집 중심에 배치
   if (old_Z=new_Z) break;
 // 과거의 군집 중심과 비교
   for (j=1 to k)
 // z_j에 배정된 샘플의 평균으로 z_j대치
 }
end:
```

Ⅳ. 유사 블로그 추천 시스템 구현

유사 블로그 추천 시스템은 TCA k-means 알고리즘을 이용하여 유사 블로그를 군집하고, 찾고자 하는 특정 주제에 맞게 관련있는 블로그를 유사도에 따라 랭크하는 시스템이다. 본 시스템은 블로그 군집 모듈과 블로그 검색 모듈로 구성되어 있다.

1. 시스템 구성도

첫째, 블로그 군집 모듈에서는 인터넷 상에 있는 블로그를 수집하여, 유사도에 따라 군집하여 DB에 저장하는 모듈이며, 둘째, 유사 블로그 검색 모듈은 블로그 군집 모듈에 의해서 군집화된 블로그들의 집단에서 찾고자 하는 주제에 해당되는 군집을 찾아 해당 주제와 관련있는 블로그를 랭크해 주는 모듈로 구성된 유사 블로그 추천 시스템의 구성도이다.

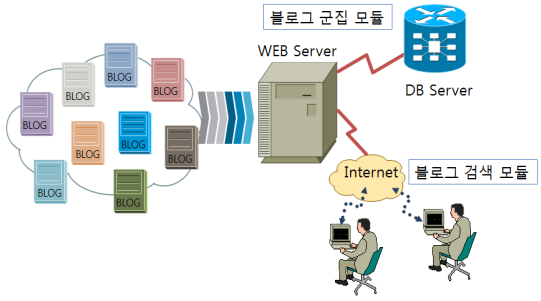


그림 2. 유사 블로그 추천 시스템
Fig. 2. System of Similar System

2. 시스템 모듈별 절차

블로그 수집 모듈은 인터넷에 흩어진 블로그를 RSS 블로그 피드 파서(blog feed parser)를 통해서 단어별로 수집하여 이차원 형태로 데이터를 압축한 뒤 TCA 군집화하여 군집 DB에 저장한다.

유사 블로그 검색 모듈은 검색하고자 하는 단어를 입력 받아서 해당 단어에 가장 유사한 블로그를 찾고, 해당 블로그에 소속된 군집 그룹을 검색하여 그룹내의 블로그와 거리를 측정한 뒤 유사도에 따라서 랭크 리스트를 화면에 출력한다.

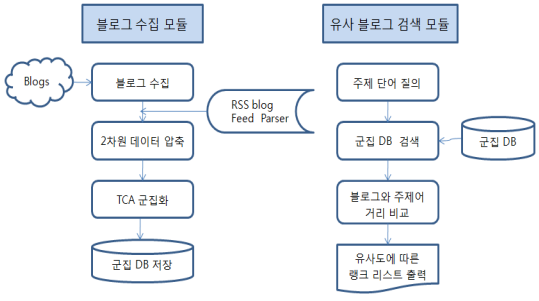


그림 3. 유사 블로그 추천 시스템 모듈별 절차
Fig. 3. A Module Procedure of similar Blog Recommandation

3. 블로그 군집 모듈

블로그 군집 모듈은 인터넷 상에 있는 블로그를 RSS 피드용 리스트 형태로 수집하여 각 블로그를 피드용 파서를 이용하여 구문을 분석한다. 분석된 단어별로 고빈도와 최소 빈도를 제외하고 블로그 마다 단어수를 세어 블로그별 단어 목록을 만든다.

(1) 블로그 수집

블로그 수집을 위해서 kiwitobes.com에서 제공하는 RSS 피드용 블로그 리스트를 담고 있는 feedlist.txt[8,9] 파일을 이용하였다. feedlist.txt에서는 99개의 블로그 URL리스트를 담고 있다.

표 3. 피드용 블로그 리스트 목록
Table 3. Feedback blog list

http://feeds.feedburner.com/37signals/beMH http://feeds.feedburner.com/blogspot/bRuz http://battellemedia.com/index.xml http://blog.guykawasaki.com/index.rdf http://blog.outer-court.com/rss.xml http://feeds.searchenginewatch.com/sewblog http://gizmodo.com/index.xml http://gofugyourself.typepad.com/go_fug_yourself/index.rdf http://googleblog.blogspot.com/rss.xml http://feeds.feedburner.com/GoogleOperatingSystem http://headrush.typepad.com/creating_passionate_ users/index.rdf ...

표 4. 수집된 단어 목록
Table 4. Collected word list

china kids music yahoo want wrong service tech saying lots had address working following years didn internet wants photos former technology being traffic small past full November experience door company learn paper research sell self sometimes couple video makes next process books could stuff audio web become problem details worth provide feeds another john away hand thanks night test update guy cost product still non drop year tried america amp start podcast month advertising ask almost ...

유니버설 피드 파서[9]를 이용하여 블로그들을 파싱하고, 단어별로 빈도를 측정한다. 모든 단어를 수용하지 않고, 저빈도와 고빈도 단어는 제외시켰다. 빈도율이 10%~50%이내의 단어만 수용하여 수집결과 99개의 블로그에서 706개 단어로 수집되었다.

(2) TCA 군집화

TCA 군집화를 통해서 영역별로 같은 모양으로 군집된 화면이다. 각 군집 그룹의 중심점을 (+)로 표시 하였으며, 영역별로 고르게 군집된 모습을 볼 수 있다.

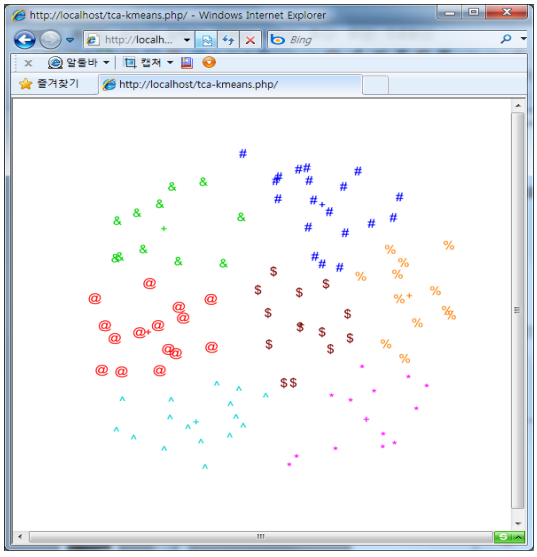


그림 4. TCA 군집 화면
Fig. 4. Termite colony algorithm Clustering Monitor

(3) 유사 블로그 추천 시스템 구현 화면

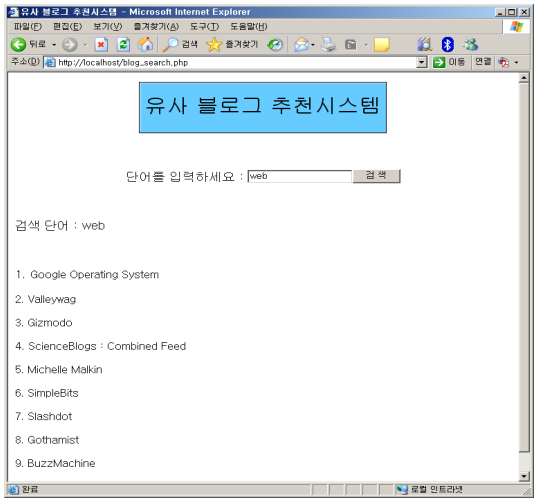


그림 5. 시스템 화면
Fig. 5. System screen

구현 환경은 Intel 2.2Ghz, RAM 2GB, Apache 2.2.14, PHP 5.3.0, Mysql 5.1.39, Windows XP, APMSETUP7, CUBRID/JRE 배포판에서 이루어졌다. [그림 5]은 HTML 입력 폼을 통해서 질의 단어를 입력받아서 출력된 유사 블로그 랭크 리스트 화면이다.

4. K-means와 TCA k-means 비교 평가

가장 널리 사용되는 군집 기법으로 K-means 알고리즘이 있다. 이 알고리즘과 본 논문에서 제안한 TCA 적용 k-means를 비교해 본 결과 군집 형태에 따른 제곱오류가 적었으며 군집 이동횟수가 줄고 평균 수행시간도 단축되었다.

실험평가의 시스템 환경은 Intel Core2 Duo 2.2Ghz, RAM 2GB, Apache 2.2.14, PHP 5.3.0, Windows XP, APMSETUP7, CUBRID/ JRE 배포판에서 이루어졌다.

[표 5]는 무작위로 주어진 99개의 샘플을 기준으로 반복 횟수에 따라서 반복 실행한 결과이다. 제곱오류(squard error) 최대값 비교에서 보면 TCA k-means의 제곱오류값이 작게 나타났다. 이는, 다양한 군집형태에 영향이 적음을 뜻한다고 볼 수 있다.

[표 6]은 반복 횟수에 따른 군집 이동횟수를 나타낸 것으로, 반복 횟수가 많을수록 이동횟수가 늘어나는 것은 무작위로 주어지는 샘플의 복잡성이 반복하는 횟수 만큼 여러 차례 나왔기 때문이다.

2000번의 반복 누적 수행시간은 k-means의 경우 29.41초, TCA k-means는 19.92초가 소요되었다. 추세 흐름으로 볼 때, 샘플 데이터 양이 많을수록 k-means 수행시간은 기하 급수적으로 늘어날 것이다.

표 5. 반복 횟수에 따른 제곱오류값 비교
Table 5. Compare Squared error values according to the number of iterations

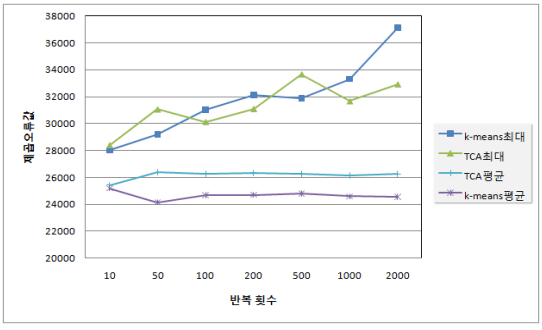
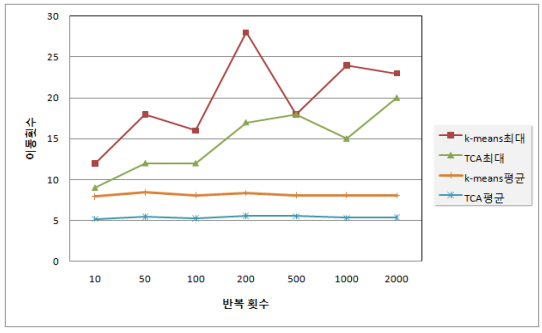


표 6. 반복 횟수에 따른 이동횟수 비교
Table 6. Comparison of the number of movements according to the number of iterations



5. 군집결과

그림 6의 군집결과를 보면, 각 색깔별로 군집된 블로그의 이름들이 나열되어 있다. 각 영역별로 골고로 군집된 형태를 나타낸다.

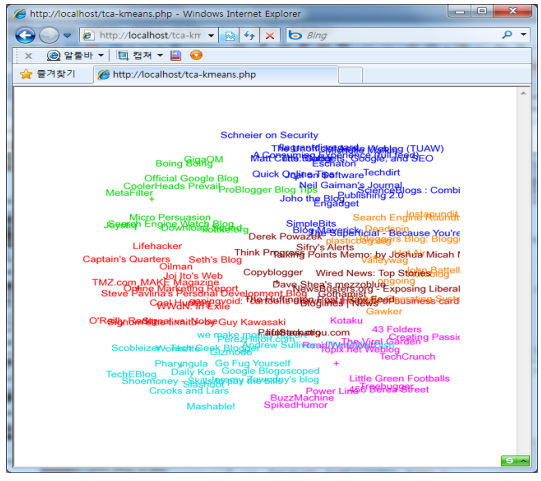


그림 6. 군집된 화면
Fig. 6. Result of Clustering

V. 결과

본 논문에서는 검색 주제어에 유사한 특성을 갖고 있는 블로그를 찾아서 유사도에 따른 추천 시스템을 제안하였다. 제안된 시스템은 블로그가 갖고 있는 특성단어를 각각 비교하여 유사도에 따라 군집화하여 검색하고자 하는 블로그의 특성 그룹을 랭크하는 추천 시스템이다.

보다 나은 군집을 위해 흰개미 습성을 이용한 TCA k-means 알고리즘을 제안하였다. 제안된 알고리즘을 실험평가한 결과 다음과 같은 논문의 성과를 얻을 수 있었다. 첫째, 군집 성능을 좌우하는 제곱오류율이 복잡한 형태에서 개선되었으며, 둘째, 군집점 이동횟수가 줄어들었고 셋째, 군집 수행시간이 단축되었다.

참 고 문 헌

[1] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, Vol.16, No.3, pp.645-678, May 2010.

[2] Su, N.-Y., Stith, B.M., Puche, H., Bardunias, P., 2010. Characterization of Tunneling Geometry of Subterranean Termites (Isoptera: Rhinotermitidae) by Computer Simulation. Sociobiology 44 (3), 471 - 483.

[3] Lee, S.-H., Bardunias, P., Su, N.-Y., 2009. Food Encounter Rates of Simulated Termite Tunnels with Variable Food Size / Distribution Pattern and Tunnel Branch Length. J. Theor. Biol. 243, 493 - 500.

[4] S.-H. Lee, N.-Y. Su and P. Bardunias, 2007. Optimal Length Distribution of Termite Tunnel Branches for Efficient Food Search and Resource Transportation, Biosystems 90 (2007), pp. 802 - 807

[5] http://kiwitobes.com/clusters/feedlist.txt

[6] http://www.feedparser.org

[7] Toby Segaran, Programming Collective Intelligence, p.77-80, O'Reilly, 2007.

[8] http://ko.wikipedia.org, "metablog"

[9] http://www.bbakorea.org

※ 본 논문은 2010년도 원광대학교 교비 지원에 의해서 수행되었으며 이에 감사드립니다.

저자 소개

정 기 성(회원)

• 전북대학교 항정학박사
• 원광대학교 소방행정학부 교수.

조 이 석(회원)

• 2007년 2월 : 군산대학교 수리정보통계학과 졸업
• 2009년 8월 : 전북대학교 컴퓨터공학과 석사

<주관심분야 : 생태모델링, 컴퓨터공학, 모바일>

이 말 례(정회원)

• 1998년 2월 : 중앙대학교 박사
• 1999년 6월 : 전남대학교 멀티미디어공학과 조교수
• 2003년 6월 : 전북대학교 컴퓨터공학과 교수

<주관심분야 : 인공지능, 로봇, 헬스케어, 게임, 가상현실 등>