Journal of The Korea Society of Computer and Information
Vol. 20, No. 2, February 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.2.029

국내 휴대폰의 진화패턴 규명을 위한 텍스트 마이닝 방안 제안 및 사례 연구

온 병 원 *

A Case Study of a Text Mining Method for Discovering Evolutionary Patterns of Mobile Phone in Korea

Byung-Won On*

요 약

생물의 진화패턴과 원리는 지난 200년간 학문적인 영역에서 활발히 연구되어 왔으며 생명의 진화에 대한 체계적인 이론, 개념 및 방법론이 제시되었다. 그리고 진화경제학, 진화심리학, 진화언어학 등 다양한 분야에 적용되어 큰 연구 성과를 거두고 있다. 이와 더불어 진화생물학 논리를 인간이 만든 제품에 적용하려는 시도도 병행되어 왔다. 기존 연구들이 생물진화 논리를 인공물에 그대로 적용하거나 해당 분야 전문가의 직관에 근거하여 진화 모형을 구축하는 것이어서 진화 모형에 대한 일반화를 시키기에는 한계를 가진다. 또한 생물과 달리 인공물은 인간 의지의 상상력이 반영되기 때문에 생물진화 이론을 곧바로 적용할 수 없다고 알려져 왔다. 따라서 본 논문에서는 특정인의 주관에 벗어나 일반 대중들의 생각을 엿보고 이를 바탕으로 진화 모형을 구축하는 것을 목표로 한다. 이를 위해, 인공물을 계통적으로 분류할 수 있는 체계적인 틀을 제시하는 텍스트 마이닝 방안과 그 결과물을 효과적으로 보여줄 수 있는 시각화 방안을 차례로 제안한다. 특히, 제안방안을 바탕으로 최근 혁신의 아이콘으로 떠오르고 있는 휴대폰과 스마트폰에 대한 사례 연구를 집중적으로 수행한다. 지난 10년간 국내에서 출시된 휴대폰과 스마트폰에 대한 리뷰 포스트들을 수집하고 분석하여, 진화패턴을 발견하고 요약해서 보여주며 그 결과에 대해서 자세히 토의한다. 더욱이 이러한 작업은 소수의 전문가들이 방대한 문헌과 자료를 조사 정리하여, 오랜 시간에 걸쳐 진화계통도를 그리게 되는 매우 지난한 작업이다. 하지만 본 논문에서 제안한 방안은 반자동(semi-automatic) 마이닝 알고리즘으로 인간의 노력을 최소화할 수 있어 그 효용 가치가 높다. 이러한 연구를 통해 인간의 창의력과 상상력이 구현되는 방식을 이해하고 휴대폰의 미래 모습을 전망하는데 있어 유관기업들에게 큰 도움을 줄 것이다.

▸ Keywords : 휴대폰, 진화패턴, 텍스트 마이닝, 비주얼라이제이션

•제1저자 : 온병원 •교신저자 : 온병원
•투고일 : 2014. 9. 12, 심사일 : 2014. 10. 7, 게재확정일 : 2014. 12. 6.
* 군산대학교 통계컴퓨터과학과(Dept. of Statistics and Computer Science, Kunsan National University)
※ 이 논문은 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2014S1A5B6037845)

Abstract

Systematic theory, concepts, and methodology for the biological evolution have been developed while patterns and principles of the evolution have been actively studied in the past 200 years. Furthermore, they are applied to various fields such as evolutionary economics, evolutionary psychology, evolutionary linguistics, making significant progress in research. In addition, existing studies have applied main biological evolutionary models to artifacts although such methods do not fit to them. These models are also limited to generalize evolutionary patterns of artifacts because they are designed in terms of a subjective point of view of experts who know well about the artifacts. Unlike biological organisms, because artifacts are likely to reflect the imagination of the human will, it is known that the theory of biological evolution cannot be directly applied to artifacts. In this paper, beyond the individual's subjective, the aim of our research is to present evolutionary patterns of a given artifact based on peeping the idea of the public. For this, we propose a text mining approach that presents a systematic framework that can find out the evolutionary patterns of a given artifact and then visualize effectively. In particular, based on our proposal, we focus mainly on a case study of mobile phone that has emerged as an icon of innovation in recent years. We collect and analyze review posts on mobile phone available in the domestic market over the past decade, and discuss the detailed results about evolutionary patterns of the mobile phone. Moreover, this kind of task is a tedious work over a long period of time because a small number of experts carry out an extensive literature survey and summarize a huge number of materials to finally draw a diagram of evolutionary patterns of the mobile phone. However, in this work, to minimize the human efforts, we present a semi-automatic mining algorithm, and through this research we can understand how human creativity and imagination are implemented. In addition, it is a big help to predict the future trend of mobile phone in business and industries.

▸ Keywords : Mobile phone, Evolutionary pattern, Text mining, Visualization

I. 서 론

2007년 애플의 아이폰이 등장하면서 현대인의 삶을 크게 바꾸어 놓았다. 스마트폰을 이용하여 언제, 어디서든지 필요한 정보를 빠르게 얻을 수 있을 뿐 아니라 물건을 구매하고 결제할 수 있고, 원격으로 가전제품과 자동차를 제어하며, 이북(e-Book)을 다운로드 받아 독서를 즐길 수 있다. 급기야 이제는 스마트폰 없는 세상은 상상할 수 없을 정도가 되었다. 스마트폰, 3D 프린터와 같은 혁신적인 상품을 통해 기업은 미래의 경쟁력을 확보하고 지속적인 성장을 하는 것이 무엇보다 중요하게 되었다. 새로운 제품군의 시장을 개척하고 선점하기 위해서는 제품의 디자인, 기능 및 사용된 기술, 그리고 아이디어 생성 메커니즘에 대한 심도 있는 이해가 필요하다. 초기에는 신제품에 대한 다양한 아이디어들이 제시되지만, 결과적으로 소수의 아이디어가 상업적으로 성공을 거두고, 채택되지 못한 아이디어들은 틈새상품(niche product)으로 전락되거나, 시장에서 퇴출되게 된다. 일례로 최근에 출시된 스마트폰들은 정전식 터치스크린과 고속 무선 데이터 통신 및 다양한 앱을 활용할 수 있는 기능들로 인해 대중에게 큰 각광을 받고 있다. 이러한 패러다임의 변화는 물건, 제품, 도구 등으로 표현될 수 있는 인공물(artifact)의 변천에 대한 진화적 접근을 통해, 인공물의 진화에 관한 일반화할 수 있는 이론,

개념적 구성요소 및 실증적 분석방법론에 대한 연구의 필요성이 증대되었다.

학계에서는 오래 전부터 진화생물학적 논리를 인공물에 적용하려는 시도는 있어 왔으나, 아직 체계적으로 정립되지 못하고 있는 실정이다(그림 1 참조). Butler와 Rivers는 인공물 역시 생물과 유사한 논리에 따라 진화한다는 주장을 제기하였으며, 주로 디자인 관점에서 인공물의 진화 논의를 전개하였다[19]. 실제로, 생물이 DNA, 환경적합, 돌연변이 속성에 의해 진화가 발생한다면, 인공물은 유사하게 핵심속성, 시장적합, 기술혁신을 통해 변천하는 것을 알 수 있다. 그러나 한편으로는 생물과 인공물의 큰 차이가 있기 때문에, 생물진화 논리를 인공물 진화에 그대로 적용할 수 없다. 예를 들면, 생물 진화가 가치중립적, 종간 유전적 소통 불가, 자연선택의 특징을 보인 반면, 인공물 진화는 가치 지향적, 종간 유전적 소통가능, 용불용설 등의 다른 특징을 보인다. 이러한 차이점과 유사점을 고려하여 인간이 만든 인공물이 진화하는 논리의 핵심 구조를 파악하는 것이 필요하다. 즉, 인공물의 진화계통도를 구축하고, 역사적 변천을 추적함으로써, 인간의 상상력이 사물로 구현되는 방식, 나아가 이러한 상상력과 인공물이 전수, 전파, 변형되는 과정에 대한 체계적인 이해를 도모할 수 있다. 뿐만 아니라, 기업의 연구개발, 신제품 개발과 경쟁전략, 정부 혁신정책 및 산업정책 수립의 학술적 근거를 제공할 수 있다[9, 10].

본 논문에서는 지난 10년간 국내에서 출시된 휴대폰과 스마트폰의 진화패턴과 그 원인을 규명하기 위한 일환으로 실증적 분석방법론을 제시하고, 사례 연구를 통해 휴대폰의 진화패턴과 원인을 규명하고자 한다. 특히, 휴대폰의 변천과정을 추적하는 실증적 분석방법론으로 휴대폰 리뷰 사이트들에 게재된 리뷰 포스트들을 수집하고 분석하여, 진화패턴을 발견하고 요약해서 보여주는 텍스트 마이닝과 시각화 방안을 제안한다. 기존 연구들이 단순히 생물진화 논리를 인공물에 적용하거나[17], 해당 분야 전문가의 직관에 근거하여 휴대폰 진화 모형을 구축하여[20], 일반화에 한계를 가지고 있었다. 본 연구에서는 수많은 대중이 이야기하고 있는 내용을 엿보고 이를 토대로 진화 모형을 구축하기 때문에, 특정인의 주관적인 의견이 반영되는 것을 피할 수 있으며 인공물 진화계통도를 그리는데, 인간의 노력을 최소화하는 반자동 알고리즘을 제안한다.

그림 1. 헬멧의 진화계통도[13]
Fig. 1. Phylogeny of Helmet

II. 관련 연구

1. 인공물 진화패턴 규명에 관한 연구

인공물 진화에 대한 연구는 디자인, 과학기술사학, 산업공학, 기술경영, 진화경제학, 컴퓨터공학 등 다양한 분야에서 시도되고 있다. 먼저 산업 디자인에서는 인공물을 규정하는 디자인 특징을 추출하고, 이를 토대로 시대적으로 인공물의 변천을 나열하는 연구가 진행되었다[20]. 과학기술사학 분야에서는 기존 인공물이 기대를 충족시키지 못하고 대안이 만들어지는데, 정치적, 경제적, 사회적, 문화적인 요인에 의해 종국에는 특정 대안만이 선택된다는 논리를 규명하는데 초점을 맞추어 왔다[8, 14]. 산업공학에서는 주로 특허를 분석하여 기술 로드맵을 만들고 미래에 대한 전망을 예측하는 연구를 수행하고 있다. 인공물 진화 메커니즘을 설명하기 위해, Arthur는 기술 사이클 이론을 정립하고 네트워크 분석법을 이용하였다[1]. 또한 진화경제학에서는 경제적 변화 과정을 다양성의 창출과 선택 과정으로 바라보고, 진화시뮬레이션 모형을 만들기 위해 활발한 연구가 진행되고 있다[7, 23]. 이밖에 Protégé & SemanticWorks(2006)은 온톨로지를 이용하여 인공물의 변천과정을 표현하는 방식을 새롭게 제시하였다.

한편 컴퓨터를 활용하여 인공물의 진화패턴을 규명하려는 연구들이 시도되었다. Best는 USENET 뉴스기사를 가지고, 뉴스간의 유사도(similarity)를 측정하고 이를 기반으로 뉴스기사들을 클러스터링 하였다[3]. 이때 뉴스간의 유사도를 밈(meme)으로 정의하였고 각 클러스터를 종(quasi-species)로 규정하고, 종간의 상호작용(interaction)을 규명하는데 초점을 맞추었다. 또한 종간의 시계열 상관관계

(cross-correlation)을 계산하여, 종간에 낮은 수준의 상관관계가 존재함을 보였다. 이와 같이 Best는 종간의 상호작용을 파악하여 인공물의 돌연변이(mutualism), 경쟁(competition), 중립(neutralism), 포식자(predator)와 피식자(prey) 관계를 규명하려는 첫 시도였다. 또한 Khanafiah와 Situngkir는 휴대폰의 스펙을 분석하여 특징 세트(feature set)을 정의하였다. 예를 들면, 특징 세트 $F=\{f_1,f_2,f_3\}$라고 하면, 휴대폰 $phone_1=\{1,0,1\}$로 표현할 수 있으며, $phone_1$은 f_1과 f_3의 특징을 가지고 있다고 할 수 있다. 이와 같이 휴대폰을 이진 벡터로 표현하고, 벡터 간의 유사도를 측정하여 휴대폰간의 거리(distance)를 계산한다. 그리고 비가중산술평균집단비교법(Unweight Pair Group Method with Arithmetic Mean)과 최단수형도(Shortest Tree)알고리즘을 사용하여, 노키아에서 만든 휴대폰 디자인을 분석한 후에 Phylomemetic Tree를 생성하였다. Khanafiah와 Situngkir의 연구는 생물학에서 흔히 사용되는 진화계통도를 인공물에 적용한 첫 사례이지만, 여러 문제점을 가지고 있다. 시간 순서가 고려되어 있지 않고, 사용한 데이터의 규모가 작을 뿐 아니라 오직 휴대폰의 디자인만을 고려하여 휴대폰의 진화 패턴을 규명하고 있다.

2. 텍스트 마이닝에 관한 연구

Blei et al는 확률 기법을 기반으로 문서의 토픽을 파악할 수 있는 토픽 모델링 알고리즘 중의 하나인 Latent Dirichlet Allocation (LDA)를 제안하였다[4]. 각 문서에는 여러 토픽들이 섞여 있고 서로 다른 확률로 분포되어있다는 가정 하에 결합확률분포를 계산하여 토픽 키워드를 추출한다. 나아가 Blei et al은 문서의 순서를 고려하여 시간의 흐름에 따른 토픽의 변화를 추적하는 연구를 수행하였다[5]. 한편 Rosen-Zvi et al는 LDA를 통해 다중 저자를 가진 논문에서 어떤 부분을 어떤 저자가 담당하는지에 대한 연구를 수행하였다[25]. 또한 이미지, 유전 정보, 통계 수치 등 다양한 데이터에서도 토픽을 추출하는 연구가 진행되어 왔다[12].

또한 국내 주요 언론사의 논조 차이점을 텍스트 마이닝 기법을 통해 알아내는 연구가 수행되었다[15]. 예를 들면, 동아일보, 경향일보 등의 언론사에서 특정 사건과 관련 있는 키워드의 빈도수를 계산하거나 코사인 유사도를 사용하여 키워드 간의 상관관계를 나타내는 그래프를 만들고 클러스터링 기법을 사용하여 패턴을 추출하였다. 현재 뉴스, 블로그와 같은 시간 정보를 포함하고 있는 문서들로부터 토픽 탐지 및 추적을 자동화하는 트렌드 분석이 활발히 진행되고 있다. 대부분의 연구들이 트렌드와 관련된 단어의 출현 빈도 정보를 이용하여 시간에 따른 트렌드 그래프를 보여주고 있다. 이러한 트렌드 곡선으로부터 변동성, 지속성, 안정성, 누적량 등의 네 가지 속성을 정의하고, 이들을 조합하여 트렌드의 순위를 결정하는 새로운 함수를 제안하였다[24]. Mei와 Zhai는 뉴스 기사로부터 주제(theme)를 추출하고, 주제들의 변화 상태를 시간 순서대로 추적하여, 특정 주제의 라이프 사이클을 파악하고, 특정 주제가 소멸되는지 또는 주제가 점점 화제가 되는지의 이유성을 찾아내는 알고리즘을 개발하였다. EM 알고리즘을 사용하여 텍스트 마이닝을 수행하였고, Hidden Markov Model (HMM) 기반의 알고리즘을 제안하였다[21].

Sun et al은 텀(term)과 세그먼트(segmentation)의 상호 정보량(mutual information)을 기반으로 문서들을 토픽별로 분절화(segmentation)시키고, 유사한 세크먼트들을 그룹핑 하는 연구를 수행하였다[26]. 또 다른 세그멘테이션 방법으로는 Misra et al은 토픽과 세그멘테이션의 결합분포를 계산하여 각 세그멘테이션의 주제 내용에 대한 정보를 모으고 이것으로부터 토픽을 알아내는 모델을 제안하였다[22]. Dalvi et al은 야후의 지역 데이터베이스를 연결시키는 생성확률모델을 제안하였다[11]. 이 연구에서는 특정 식당에 대한 리뷰는 특정 식당과 관련된 단어들로 구성되고, 그 식당과 관련 없는 단어들은 그 리뷰에 포함되지 않는다는 Generic Review Language (RLM) 모델을 기반으로 한다. 그리고 특정 식당을 질의, 리뷰를 문서로 가정하여 RLM 모델을 사용할 경우에 기존의 TF/IDF 매칭 알고리즘을 사용하는 방법에 비해 높은 정확도를 보였다.

최근 국내에서는 텍스트 마이닝 기법을 이용한 다양한 적용 사례들이 나타나고 있다. 다음소프트에서는 소셜 미디어 데이터를 자연어 처리 기술로 분석하였다. 한국전자통신연구원이 개발 중인 소셜 미디어 이슈 탐지 및 모니터링 플랫폼인 WISDOM은 소셜 미디어를 수집, 저장하고 심층적인 언어분석을 기반으로 추출한 정보를 이용하여 이슈를 탐지하고 모니터링 하는 기능을 제공한다[18]. 이와 관련된 해외 사례로는 Recorded Future가 대표적이다. 웹 사이트, 블로그, 소셜 미디어 등의 구조화되지 않은 대규모의 텍스트 데이터를 대상으로 정보를 추출, 분석, 시각화 서비스를 제공하고 있다[28].

III. 제안 방안

본 연구는 지난 10년 동안 국내에서 출시되었던 휴대폰과 스마트폰의 진화패턴을 규명하기 위한 방안을 새롭게 제안한다.

먼저 휴대폰의 진화를 설명하기 위해서는 생물 진화에서 핵심이 되는 유전자(gene)와 운반자(vehicle) 개념을 적용시킬 필요가 있다. 모든 생물체는 진핵세포에 흔히 유전자라 부르는 DNA 염기서열을 가지고 있고, A-C-G-T와 같은 4가지 염기가 중합되어 이중나선 구조를 이룬다. 이러한 유전 정보를 바탕으로 뇌, 심장, 근육 등이 만들어진다. 이때, 유전자에 의해 만들어진 발현물(發現物)을 운반자라 할 수 있다. 예를 들면, 어떤 특정 유전자를 통해 벽안(碧眼)이 만들어진다고 하면, 벽안은 운반자가 된다. 유전자는 진화의 과정에서 자기 자신을 더 많이 퍼뜨리도록 행동하고 그러한 유전자만이 살아남는다. 본 논문에서는 인공물, 특히 휴대폰에서 운반자는 '휴대폰의 특징'이라고 가정한다.[1] 또한 유전자는 휴대폰의 특징을 정의하는 틀 혹은 설계도라고 정의하며, '속성(property)'과 '기능(function)'으로 구성된다. 예를 들면, DMB는 휴대폰의 한 특징이고, 유전자는 작동(on/off), 채널(channel), 음량(volume) 등과 같은 속성(상태)과 켜기(power on), 끄기(power off), 채널 증가(increase channel), 채널 감소(decrease channel), 음량 증가(increase volume), 음량 감소(decrease volume) 등과 같은 기능(행위)을 가질 수 있다. 이러한 유전자는 운반자(휴대폰의 한 특징 - DMB)를 조정함으로써 간접적으로 환경(시장)과 상호작용하게 된다. 그리고 1859년 다윈이 종의 기원(On the origin of species of natural selection)에서 밝혔던 것처럼 변이(variation), 적합도(differential fitness), 유전(heredity)라는 3가지 필요충분조건에 의해 생물체의 진화가 발생된다고 하였다. 마찬가지로 인공물에서도 기술의 참신함을 변이로 볼 수 있고, 조상관계로 맺어진 인공물들의 핵심적인 공통 특징이 유전되는 것을 볼 수 있다. 또한 인공물은 대중의 선택에 의해 진화가 발생하는 것이 보편적이며, 이 때 대중의 선택은 시장이나 생물체 진화의 환경인 자연이라고 은유할 수 있다. 기능적으로 매우 우수한 제품이 출시되었다 하더라도, 어떤 이유에서건 대중의 외면을 받는다면(환경에 적응하지 못한다면), 그 제품은 더 이상 존재할 수 없다. 이런 측면에서 진화를 일으키는 다양한 원인 중에서 사용자의 선택(user preference)이 인공물의 진화를 일으키는데 있어 주요한 동인이라 할 수 있다[2]. 특히, 어떤 제품에 대한 사용자의 선택을 쉽게 알 수 있는 바로미터로는 리뷰 사이트(review site)를 꼽을 수 있다. 사용자들은 리뷰 사이트에서 제품에 대한 평가를 한 후에 개인적인 의견을 반영하는 리뷰 포스트(review post)를 수시로 올리고 서로 의견을 교환한다. 이와 같이 다수의 리뷰 사이트에서 대용량 텍스트 리뷰 포스트들을 수집하고 분석하여 진화패턴을 규명하는 것이 필요하다. 그리고 이러한 작업의 핵심은 텍스트 마이닝 기법을 사용하여 리뷰 포스트들을 분석하고 자동으로 요약해주는 알고리즘의 개발이 필요하다. 특히, 텍스트 마이닝을 통해 추출된 진화패턴을 요약하는 방법으로 그래프 시각화 방안을 제안한다. 또한 지난 10년간 국내에 출시된 주요 휴대폰을 대상으로 사례 연구(case study)를 진행하고 진화패턴에 대해 자세히 논의한다. 분석 결과를 토대로, 휴대폰 또한 자연선택의 3가지 필요충분조건인 변이, 적합도, 유전에 의해 진화가 일어나는지를 규명하고 생물체와 휴대폰의 진화가 어떻게 다르며, 휴대폰이라는 인공물 진화에 필요한 독특한 진화 원리가 있는지를 토의한다.

Algorithm 1은 세티즌(Cetizen)과 오픈모바일(Open Mobile)에서 수집한 리뷰 포스트들을 분석하여 추출한 진화패턴을 그래프로 시각화하여 보여주는 알고리즘을 나타낸다.

Algorithm 1: 휴대폰의 진화패턴 추출 알고리즘
Input: 리뷰 포스트 컬렉션(a collection of review posts)

1단계: 원자료 전처리
  Output: 분기별 리뷰 포스트 그룹
2단계: 분기별 리뷰 포스트 그룹의 텍스트 요약
  Output: 분기별 리뷰 포스트 그룹의 토픽 추출
3단계: 토픽으로부터 주요 특징(feature) 추출
  Output: 분기별 휴대폰의 특징 추출
4단계: 주요 특징을 휴대폰에 매칭
  Output: 분기별 〈휴대폰1, 특징1, 특징2, …〉, 〈휴대폰2, 특징1, 특징3,…〉, …
5단계: 진화계통도를 위한 그래프 생성
  Output: 시계열 그래프 (노드: 휴대폰, 링크: 특징 상속)

제안방안은 5단계로 구성되며, 원자료 전처리 단계에서는 국내의 리뷰 사이트에서 리뷰 포스트들을 수집하고 정제한 후에 데이터베이스에 저장한다. 2단계에서는 분기별로 그룹 된 리뷰 포스트들을 분석하여 그 당시 사용자간에 활발히 화자되었던 토픽(topic)들을 추출한다. 3 단계에서는 추출된 토픽들을 분석하여 분기별 휴대폰의 주요 특징(feature)을 추출한다. 그리고 다음 단계에서 분기별로 출시되었던 휴대폰과 이미 추출된 주요 특징들을 매칭 함으로써, 각 휴대폰의 주요 특징들을 파악한다. 마지막 단계에서는 진화계통도를 위한 그래프를 생성한다. 각 휴대폰은 그래프의 노드로 나타내고, 과거와 현재에 출시된 휴대폰간의 유사한 특징들이 존재하면, 과거에 출시된 휴대폰으로부터 현재 출시된 휴대폰으로 링크

1) 최근 휴대폰에는 스마트폰 여부, 디자인, 외산, 운영체제 종류, 디스플레이 인치, 터치 기능, 내장메모리, LTE, 영상통화, 휴대폰 무게, 자이로스코프센서, MP3, DMB, 카메라, 배터리 용량, AP, 앱, 멀티태스킹 지원 여부 등 다양한 특징들이 나타난다.

를 연결한다. 본 논문에서는 이러한 링크를 특징 상속(feature inheritance)라고 부른다. 이러한 과정을 반복하여 지난 10년 동안 국내에서 출시된 휴대폰과 스마트폰의 진화패턴을 규명하는 그래프를 완성한다.

다음 장에서는 제안방안의 각 단계에 대해 자세히 살펴본다.

1. 원자료 전처리

세티즌과 오픈모바일 등 국내 주요 리뷰 사이트에서 리뷰 포스트를 수집한 다음, 각 리뷰 포스트에서 제목, 날짜, 텍스트를 추출한다. 그림 5는 세티즌에 게시된 휴대폰의 리뷰 포스트에 대한 예제이다. 그리고 한글 형태소 분석을 통해, 명사 식별 및 띄어쓰기 교정 등을 거쳐 텍스트 데이터를 정제한 후에 데이터베이스에 저장한다. 본 연구에서는 국민대의 한글 형태소 분석 라이브러리 오픈 소스[16]을 사용하여, 각 리뷰 포스트 원문 텍스트를 정제하였다.

또한 데이터베이스에 저장되어 있는 리뷰 포스트들을 기간별로 그룹으로 묶는다. 국내 휴대폰 시장에서 신제품은 3개월 단위로 출시되기 때문에, 3개월 내에 게재된 리뷰 포스트들은 같은 그룹으로 묶는다. 또한 그림 5처럼, 사용자가 작성한 리뷰 포스트는 웹사이트에 게재된 시점을 명확하게 알 수 있음으로, 별도의 복잡한 과정 없이, 3개월 단위로 리뷰 포스트들을 동일 그룹으로 묶는다.

2. 리뷰 포스트 텍스트 요약

데이터 전처리 단계를 거쳐, 데이터베이스에 있는 모든 리뷰 포스트들은 3개월 단위의 그룹들로 나뉜다. 예를 들면, 2012년 1월 ~ 3월(시간 t_0)에 게재되었던 리뷰 포스트들은 r_1, r_2, r_3, r_4, r_5라 하고 2012년 4월 ~ 6월(시간 t_1)에 게재되었던 리뷰 포스트들은 $r_6, r_7, r_8, r_9, r_{10}$ 라고가정하면, $t_0 = \{r_1, r_2, r_3, r_4, r_5\}$와 $t_1 = \{r_6, r_7, r_8, r_9, r_{10}\}$로 그룹핑할 수 있다.

이 단계에서는 토픽 모델링 알고리즘(topic modeling algorithm)을 사용하여 t_0에 있는 리뷰포스트들을 클러스터링(clustering)하여 클러스터 세트(cluster set)를 생성한다. 같은 방법으로 t_1에 있는 리뷰포스트들을 클러스터링 하여 클러스터 세트를 생성한다. 예를 들어, t_0에 있는 리뷰 포스트들로부터 다음 2개의 클러스터 세트(C_1과 C_2)들을 얻었다고 가정하자. 이를 테면, $C_1 = \{r_1, r_4, r_5\}$와 $C_2 = \{r_2, r_3\}$라고 한다면, C_1에 속하는 r_1, r_4, r_5 리뷰 포스트들은 내용이 서로 유사해야 하며, C_2에 속하는 리뷰 포스트들과 다르면, 클러스터링이 효과적으로 수행되었다고 할 수 있다. 또한 C_1에 속한 리뷰 포스트들을 분석하여, 토픽(a set of topics)들과 토픽들의 확률 분포(probability distribution)를 추출할 수 있다면, r_1, r_4, r_5 리뷰 포스트들이 실제로 휴대폰의 어떤 내용을 다루고 있는지를 파악할 수 있게 된다. 예를 들면, 해당 리뷰 포스트들에서 특정 스마트폰의 디스플레이(display)와 배터리(battery) 기능들을 다른 회사의 제품들과 비교한다면, 그 리뷰 포스트들의 주요 토픽(주제)들은 '디스플레이'와 '배터리'가 될 것이다. 또한 해당 리뷰 포스트들 중에서 디스플레이에 대해 60%, 배터리에 대해서 40% 정도의 비율로 이야기되고 있다면, 토픽들에 대한 확률 분포는 각각 0.6과 0.4임을 알 수 있다. 따라서 $t_0 = \{r_1, r_2, r_3, r_4, r_5\}$이 입력으로 주어지면, 토픽 모델링 알고리즘을 사용하여 해당 리뷰 포스트들의 토픽들을 추출할 수 있다. 또한 각 토픽은 단어(명사)들의 집합이고, 단어들은 서로 연관어 관계를 가진다. 따라서 토픽 내의 단어들을 분석하면 토픽이 실제로 무엇인지를 알 수 있다. 또한 각 단어는 확률 값을 가지고 있어, 토픽 내에 그 단어의 중요성을 쉽게 파악할 수 있다.

본 연구에서는 토픽 모델링 알고리즘을 위해 Latent Dirichlet Allocation (LDA) 방식을 사용하여 리뷰 포스트들의 토픽 세트와 확률 분포를 추출한다. LDA는 문서들의 주제를 알기 위해 원본 텍스트 내의 단어를 분석하는 통계적인 방법이다. LDA는 확률 그래프 모델 중의 하나로 Dirichlet 분포를 이용하여 텍스트 문서 내의 단어들이 어떤 특정 토픽에 포함될 확률을 계산하는 모델이다. LDA는 텍스트 문서에는 여러 토픽들이 혼합되어 있다는 가정에서 출발한다. LDA는 토픽 레이어(layer)를 통해 문서 레이어와 단어 레이어로 연결(link)된다. 이와 같은 그래프를 Tripartite Graph라고 하는데, 문서, 토픽, 단어 등 3개의 다른 형태의 노드(node)들로 구성된다. 동일 그룹의 노드들은 서로 연결되지 않고, 다른 타입의 노드들과 연결된다. 예를 들면, 문서1은 토픽1, 토픽2, 토픽3에 연결되고 링크의 웨이트(weight) 값은 확률

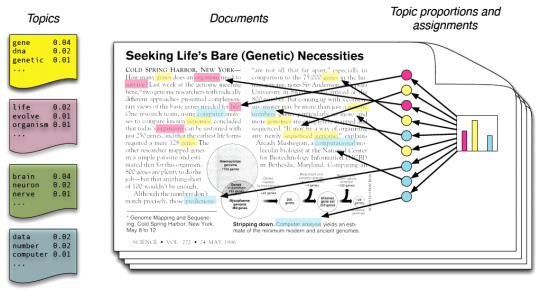


그림 2. 토픽 분포 차트[8]
Fig. 2. Topic Distribution Chart

값이다. 만일 문서1과 토픽1, 토픽2, 토픽3 각각의 확률 값이 0.6, 0.4, 0이라고 하면, 문서1은 토픽1(디스플레이)과 토픽2(배터리)로 구성된다. 그리고 '디스플레이'와 '배터리'의 확률 분포는 0.6와 0.4이다.

LDA는 또한 어떤 텍스트 문서도 생성되기 전에 이미 토픽 구조가 존재하며, 숨겨져 있는 토픽 구조 (hidden topic structure)에 의해서 문서가 생성된다는 생성확률모델(generative model)이다. 즉, 숨겨져 있는 토픽 구조를 미리 가정하고, 현재 관찰 가능한 문서 내의 단어들은 이로부터 생성되었다는 가정에서 출발한다. LDA에서 숨겨져 있는 토픽 구조(파라미터)는 다음과 같다.

1) 토픽 개수 (number of topics)
2) 문서 d의 토픽 분포도 ($\theta^{(d)}$)
3) 문서 d의 각 단어의 특정 토픽 배정 확률 ($Z^{(d)}$)
4) 토픽 Z의 단어 분포 ($\Phi^{(z)}$)

그림 2는 LDA를 설명하기 위해 사용되는 Blei의 예제 그림이다[6]. 그림에서 보면, 문서들의 집합(a collection of documents)는 크게 4개의 토픽들로 구성된다. 그림의 왼편에는 4개의 토픽들이 있는데, Genetics (노란색), Life Organism (분홍색), Brain Science (녹색), Data Analysis (파란색)에 관한 토픽들이다. 각 토픽은 단어와 단어의 확률 값으로 구성된다. 즉, Genetics 토픽에는 gene, DNA, genetic 등의 단어들과 토픽 내 단어들의 확률 분포로 구성된다. 예를 들면, 토픽 Z(예: Genetics)의 단어 분포는 $\Phi^{(z)}$이다. 그림에서 원 도형은 토픽을 나타내며, 히스토그램($\theta^{(d)}$)은 문서 별 토픽 분포도를 나타낸다. 또한 화살표는 문서 D(예: Seeking Life's Bare (Genetic) Necessities)내의 각 단어의 토픽 지정을 의미한다. 즉, 문서 d에 있는 각 단어가 특정 토픽에 있는 단어들의 확률 분포에 의해 생성됨을 의미하고, $Z^{(d)}$로 표현한다. 결론적으로, 주어진 각 문서는 본래 Genetics, Life Organism, Brain Science, Data Analysis라는 숨겨진 토픽들의 확률 분포에 의해 만들어지며, 그 문서에 있는 각 단어는 토픽 내의 그 단어의 확률 분포에 의해 문서 내에 생성된다는 것이 LDA의 핵심이다.

그림 3은 LDA의 기본 개념을 그래픽 모델(graphical model)로 도식화한 것이다. 그림에서 원 도형은 변수를 의미한다. 음영된 원 도형은 관찰 변수(observed value)를 나타내고, 흰색의 원 도형은 숨겨진 변수(hidden variable)을 가리킨다. 또한 D는 문서들의 수를 의미하고 N_d는 문서 d에 있는 단어들의 개수를 나타낸다. 문서 d와 단어 w에 대한 결합확률분포(posterior distribution)은 다음과 같이 구할 수 있다.

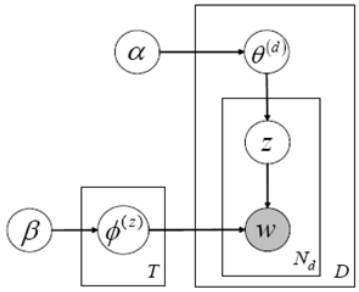


그림 3. LDA 그래픽 모델[27]
Fig. 3. Graphical Model of LDA

$$P(d,w) = P(d)P(\theta^{(d)}|\alpha)\sum_{z} P(\Phi^{(z)}|\beta)P(w|z,\Phi^{(z)})P(z|\theta^{(d)}) \quad (1)$$

수식 (1)에서 $P(d)$는 문서 d의 확률 값을 말한다. 또한 α와 β는 사전확률(prior probability)을 나타내고, α는 문서 내의 토픽 분포, β는 토픽 내의 단어가 어떤 확률로 분포하는지를 나타내는 사전확률이다. 베이지안 정리(Bayes' Rule)에 의해, 모델 M의 사후 확률(posterior probability) $P(M|x)$는 사전확률($P(M)$)과 가능도(likelihood $P(x|M)$)의 곱(product)로 구해진다. 가능도는 모집단을 추정하기 위해 모델 M으로 가정하면 (예를 들면, Gaussian distribution), 관찰 변수(x)의 확률을 구할 수 있다. 특히, 다항 분포 (multinomial distribution)에서 베이지안 모델을 고려할 경우에 Dirichlet 분포로 사전 확률을 사용하면, 사후 확률 또한 Dirichlet 분포로 알려져 있다. 따라서 LDA 모델에서는 Dirichlet prior인 α와 β을 사용한다.

그러나 수식 (1)을 사용하여 모집단(a collection of documents)의 모든 단어들을 고려하여 사후 확률을 계산하는 것은 불가능하다. 따라서 간접 방식으로 사후 확률 분포를 추정하게 된다. 사후 확률 분포를 추정하기 위해, 크게 2가지 방법이 많이 사용된다. 직접전인 방법으로는 변분법(variational method)가 있다. 이 방법은 최적화방법(optimization method)를 사용하고 복잡하기 때문에, 잘 사용되지 않았지만, 최근 들어 하둡의 맵리듀스(MapReduce) 프로그래밍이 널리 각광을 받으면서 대용량 문서를 병렬 처리하기 위해 많이 사용된다. 다른 방법으로는 간접적인 방식이 있고, 주로 깁스 샘플링(Gibbs sampling)을 사용한다. 깁스 샘플링은 2개 이상의 변수들의 결합확률분포(joint probability distribution)로부터 연속적인 표본을 샘플링을 한다. 결합분포가 명확히 알려져 있지 않으나, 각

변수의 조건부 분포는 알려져 있을 경우에 적용 가능하며, 추정하고자 하는 변수의 나머지 변수에 대한 조건부 확률분포에 의존하여 교대로 표본을 채취하는 방법으로 구현이 가능하다.

Algorithm 2: 사후 확률 추정을 위한 깁스 샘플링

Input: A collection of documents, α, β, Number of topics

1단계: Random start

2단계: For word w in document d

$$P(z_i = j|z_{-i}, w_i, d_i, \bullet) \propto \frac{c_{w_j}^{WT} + \beta}{\sum_{w=1}^{W} c_{w_j}^{WT} + w\beta} \frac{c_{d_j}^{DT} + \alpha}{\sum_{t=1}^{T} c_{d_t}^{DT} + T\alpha} \quad (2)$$

3단계: Reassign w to topic t by $P(z_i = j|z_{-i}, w_i, d_i, \bullet)$

Algorithm 2에 있는 수식 (2)의 사후 확률은 $P(t|d)$와 $P(w|t)$의 곱(product)으로 구해진다. Algorithm 2를 자세히 설명하기 위해 아래와 같은 예제를 고려한다.

- $D_1 = \{a(T_2), b(T_2), c(T_1), b(T_2)\}$
- $D_2 = \{a(T_1), b(T_2), b(T_1), d(T_2)\}$
- $D_3 = \{d(T_1), b(T_2), e(T_2), b(T_1)\}$

위의 예제에서 D_1, D_2, D_3는 각각 리뷰 포스트를 나타내며, D_1은 a,b,c,b 단어들로 구성된다. 또한 D_2은 a,b,b,d 단어들로 구성되고, D_3은 a,b,e,b 단어들로 구성된다. 그리고 Algorithm 2가 실행되기 전에 각 단어는 랜덤(random)하게 토픽 T_1이나 T_2에 어사인(assign)된다. 예를 들면, D_1의 단어 a는 랜덤하게 토픽 T_2에 어사인 되고, Algorithm 2의 1단계에 해당된다. 또한 2단계에서는 C^{WT}와 C^{DT} 테이블이 다음과 같이 만들어진다.

C^{WT}	T_1	T_2
a	1	1
b	2	4
c	1	0
d	1	1
e	0	1

C^{WT}를 통해 a 단어는 T_1 과 T_2 토픽에 각각 한 번씩 어사인 되었음을 알 수 있다. 또한 C^{DT} 테이블을 통해 D_1 문서가 T_1 과 T_2 토픽에 1번, 3번 어사인 됨을 알 수 있다.

C^{DT}	D_1	D_2	D_3
T_1	1	2	2
T_2	3	2	2

위와 같이 C^{WT}와 C^{DT} 테이블이 생성된 후에 각 단어의 확률 값을 구하고 확률 값이 큰 토픽에 다시 어사인 된다. 예를 들면, D_1의 a 단어의 확률 값을 구하기 위해, 먼저 $C^{WT}(a, T_2) = C^{WT}(a, T_2) - 1$을 하고, $C^{DT}(a, T_2) = C^{WT}(T_2, D_1) - 1$을 한다. 그리고 a 단어가 각 토픽에 속할 확률을 구한다.

$$P(Z_i = T_1|z_{-i}, a, d_i, \bullet) = \frac{1+0.01}{4+5\times0.01} \times \frac{1+25}{2+2\times25} = 0.13$$

$$P(Z_i = T_2|z_{-i}, a, d_i, \bullet) = \frac{0+0.01}{6+5\times0.01} \times \frac{2+25}{1+2\times25} = 0.00088$$

w는 단어의 개수를 말하며, T는 토픽들의 수이다. α와 β는 사전 확률로 $\alpha = \frac{50}{T} = 25$와 $\beta = 0.01$의 값을 사용한다. a 단어가 T_1 에 속할 확률이 크므로 T_1 에 어사인 한다. 마지막으로 $C^{WT}(a, T_1) = C^{WT}(a, T_1) + 1$을 하고, $C^{DT}(a, T_1) = C^{WT}(T_1, D_1) + 1$을 한다. 위와 같은 방식으로 각 단어의 토픽을 구하게 된다.

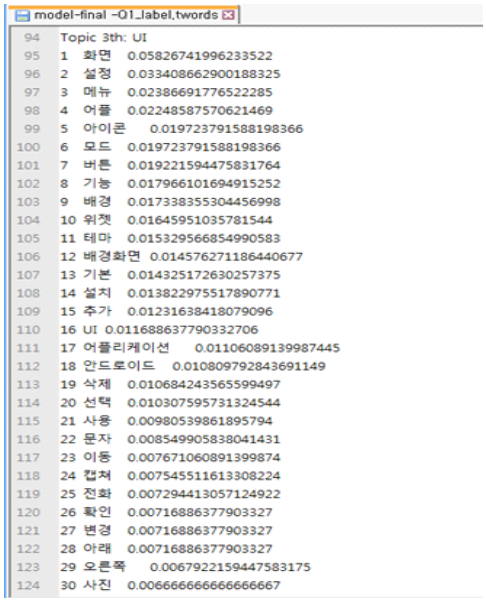
Algorithm 2을 통해 문서들의 토픽 세트가 추출된다. 예를 들면, 그림 1에서 보이는 Genetics, Life Organism, Brain Science, Data Analysis라는 4개의 토픽들이 추출된다. 각 토픽은 유사한 의미를 지니는 단어들과 그 토픽 내의 단어들의 확률 분포가 출력되어, 특정 단어가 그 토픽 내의 중요도를 판단할 수 있다. 이와 같이 Algorithm 2를 통해 추출된 토픽들로부터 다음 장에서 설명할 휴대폰의 주요 특징(feature) 추출을 위해 사용된다.

다른 토픽 모델링 알고리즘 중에서 LDA는 성능이 우수한 것으로 알려져 있다. 대량의 문서 컬렉션을 기계적으로 빠르고 정확하게 처리할 수 있어, 현재 널리 사용되고 있다. 하지만, LDA는 텍스트 문서 요약(text summarization)을 하는 통계적인 방법론이기 때문에 어떤 텍스트 도메인이 입력으로 주어지면, 단지 토픽 세트(a set of topics)만 결과로 주어진다. 각 토픽의 레이블(label)은 도메인 전문가가 토픽에 속하는 단어들을 분석하여 수작업으로 토픽의 레이블을 정해야 되는 단점이 있다. 예를 들면, 대용량 뉴스 기사로부터 LDA를 사용하여 자동으로 토픽 세트를 추출하였다고 가정하자. 각 토픽은 연관성 있는 단어들과 확률 값으로 구성된다. 그리고 도메인 전문가는 토픽을 분석하여, 각 토픽의 레이블을 결정

하게 된다. 이를 테면, 토픽 = {(실업, 0.4), (해고, 0.3), (회사, 0.3)}에서 토픽 단어들은 실업문제와 관련 있으며, 도메인 전문가는 LDA로부터 추출된 그 토픽을 '실업문제'로 레이블링(labeling)하여야 한다.

3. 휴대폰의 주요 특징 추출

III.2장에서 설명한 바와 같이, LDA를 사용하여 토픽 세트를 추출할 수 있다. 각 토픽은 유사한 의미를 지니는 단어들과 그 확률 분포로 구성되어 있다. 이러한 유사한 단어들은 휴대폰의 어떤 기능 또는 디자인을 설명하는 단어들이다. 따라서 도메인 전문가가 토픽 내의 단어들을 분석하여, 그 토픽을 레이블링 한다면, 그 토픽 레이블은 리뷰 사이트에서 휴대폰 사용자들이 게재한 리뷰 포스트에서 주로 다루어지는 휴대폰의 어떤 기능이나 디자인에 대한 것이다. 예를 들어, 토픽 = {(픽셀, 0.5), (카메라, 0.3), (화소, 0.2)}에 대해서, 휴대폰의 디지털 카메라가 토픽의 레이블이라고 예상할 수 있다. 따라서 이 토픽은 휴대폰의 한 특징(디지털 카메라)를 나타낸다고 할 수 있다.


model-final -01_label.twords

```
Topic 3th: UI
1 화면 0.05826741996233522
2 설정 0.033408662900188325
3 메뉴 0.02386691776522285
4 어플 0.02248587570621469
5 아이콘 0.019723791588198366
6 모드 0.019723791588198366
7 버튼 0.019221594475831764
8 기능 0.017966101694915252
9 배경 0.017338355304456998
10 위젯 0.01645951035781544
11 테마 0.015329566854990583
12 배경화면 0.014576271186440677
13 기본 0.014325172630257375
14 설치 0.013822975517890771
15 추가 0.01231638418079096
16 UI 0.011688637790332706
17 어플리케이션 0.01106089139987445
18 안드로이드 0.010809792843691149
19 삭제 0.010684243565599497
20 선택 0.010307595731324544
21 사용 0.00980539861895794
22 문자 0.008549905838041431
23 이동 0.007671060891399874
24 캡쳐 0.007545511613308224
25 전화 0.007294413057124922
26 확인 0.00716886377903327
27 변경 0.00716886377903327
28 아래 0.00716886377903327
29 오른쪽 0.0067922159447583175
30 사진 0.006666666666666667
```

그림 4. LDA을 사용한 토픽 결과 예제 화면
Fig. 4. An Example of Topics by LDA

그림 4는 LDA를 사용하여 추출한 한 토픽의 예제이다. 토픽에는 30여개의 단어들이 있고, 그 확률 값들이 나타난다. 또한 그 단어들이 확률 값에 의해 내림차순으로 정렬되어 있음을 알 수 있다. 본 예제에서는 '화면', '설정', '메뉴' 등의 단어가 토픽 내에서 중요한 단어들임을 알 수 있다. 휴대폰 전문가가 이러한 연관어들을 분석하여, 'UI'라고 토픽을 레이블링 할 수 있다. 이와 같은 방식으로, LDA를 사용하여 추출된 토픽 각각은 휴대폰의 한 특징이 된다고 할 수 있다. 2011년부터 2012년까지 매 분기마다 추출된 특징들을 살펴보면, 2011년 1분기에 추출된 특징들로는 '인터넷', '동영상', '자이로스코프센서', 'UI', '카메라', 'AP', '미러링', 'SNS' 등이 있다.

4. 주요 특징을 휴대폰에 매칭

각 리뷰 포스트는 대체로 하나의 특정 휴대폰에 대해서 그 특징들을 언급하기 때문에, 리뷰 포스트가 어떤 특정한 제품의 휴대폰을 이야기하는지 알 수 있다. 다시 말해, 〈리뷰 포스트, 특정 휴대폰〉의 관계를 파악할 수 있다. 이를 테면, 리뷰 포스트 r_1,r_2,r_3가 있고, r_1은 애플의 iPhone5에 대해 이야기하고, r_2는 삼성의 갤럭시S5, r_3는 LG의 G2를 이야기한다면, 〈r_1, iPhone5〉, 〈r_2, 갤럭시S5〉, 〈r_3, G2〉의 관계를 쉽게 찾을 수 있다.

만일 어떤 특징과 가장 연관성 있는 리뷰 포스트를 찾을 수 있다면, 〈특징, 리뷰 포스트〉 관계를 파악할 수 있다. 예를 들어, '디스플레이'라는 특징이 리뷰 포스트 r_1과 r_3와 연관이 있다면, 〈디스플레이, r_1〉과 〈디스플레이, r_3〉 관계를 쉽게 파악할 수 있다.

그리고 〈r_1, iPhone5〉, 〈r_2, 갤럭시S5〉, 〈r_3, G2〉와 〈디스플레이, r_1〉, 〈디스플레이, r_3〉 관계에서 이행적 폐쇄(transitive closure) 과정을 통해 최종적으로 〈디스플레이, iPhone5〉와 〈디스플레이, G2〉 관계를 얻을 수 있으며, 이것은 iPhone5와 G2는 '디스플레이'라는 주요 특징을 가진다는 것을 의미한다. 다시 말하면 리뷰 포스트를 매개로 하여 특정 시점(시간 t_0)에서 나타난 주요 특징(feature)을 특정 휴대폰(예: iPhone5)와 연결하는 과정이다.

위와 같이 III.3장에서 추출된 주요 특징을 특정 휴대폰과 짝짓기 위해서는 먼저 주요 특징과 가장 유사한 리뷰 포스트를 찾는 과정이 필요하다. 이를 위해, 본 논문에서는 매칭 알고리즘을 제안한다.

Algorithm 3: 매칭 알고리즘
Input: A collection of documents, A set of topics T, $S=0$

1단계: Segment each document to paragraphs $s_1,\ldots,s_i$
2단계: $S=S\cup\{s_1,\ldots,s_i\}$
3단계: for topic $t\in T$
for paragraph $s\in S$

Compute score(t,s) using: $\text{argmax}_s P(s|t)$
4단계: Sort paragraphs by score(s,t)s in descending order
5단계: Select top-k paragraphs

Algorithm 3에서 각 리뷰 포스트는 문단(paragraph)들로 나뉘어(segment)지고 집합 S에 저장된다. S는 문단들의 집합(a set of paragraphs)을 나타낸다. 토픽에 대한 문단의 매칭 알고리즘은 다음과 같은 가설에 의해서 동작한다.

가설: 토픽 모델을 통해 얻어진 토픽 내의 단어들과 그 확률 값들을 사용하여, 주어진 토픽에 대한 각 문단의 확률 값을 계산한다.

이러한 가설을 바탕으로, 문단 s에 대한 토픽 t의 확률 $P(t|s)$는 다음과 같이 계산할 수 있다.

$$P(t|s) = Z(t)\prod_{w\in t}P(w|s) \approx Z(t)\prod_{w\in t}P_s(w) = \prod_{w\in t}\frac{g(w)}{\sum_{w'\in t\cap text(s)} g(w')} \quad (3)$$

수식 (3)에서 t, w, s는 각각 토픽, 토픽 내의 단어, 문단 등을 의미한다. $Z(t)$는 t의 총 단어 개수로 노멀라이즈 파라미터(normalized parameter)이다. 수식 (3)를 구하기 위해서, s에 대한 토픽 내의 단어 w의 확률은 독립확률변수이기 때문에 각 단어의 확률 값의 곱으로 계산된다.

(3)의 생성확률모델을 바탕으로 하여 토픽에 대한 문단 s의 확률 값은 다음과 같이 구할 수 있다.

$$s^* = \text{argmax}_s P(s|t) \quad (4)$$
$$= \text{argmax}_s \frac{P(t|s)P(s)}{P(t)} \quad (5)$$
$$= \text{argmax}_s \frac{P(s)}{P(t)}P(t|s) \quad (6)$$
$$= \text{argmax}_s P(t|s) \quad (7)$$
$$= \text{argmax}_s Z(t)\prod_{w\in t}P(w|s) \quad (8)$$
$$= \text{argmax}_s Z(t)\prod_{w\in t}P_s(w) \quad (9)$$
$$= \text{argmax}_s \prod_{w\in t}P_s(w) \quad (10)$$
$$= \text{argmax}_s \sum_{w\in t}\log(P_s(w)) \quad (11)$$

$P(s|t)$는 베이지안 정리에 의해 (5)로 치환될 수 있다. (6)의 수식에서 $P(s)$는 알려지지 않은 값이기 때문에 t에 대해 모든 $P(s)$들을 균등분포(uniform distribution)으로 가정하여 $\frac{P(s)}{P(t)}$을 생략할 수 있다. (7)의 수식에서 $P(t|s)$는 앞에서 제안한 생성확률모델을 기반으로 하여 (3)의 수식을 사용해서 (9)로 치환할 수 있다. 또한 $Z(t)$는 s에 독립변수이기 때문에 생략한다. (10)의 수식에서 확률분포함수가 곱셈 꼴일 때 미분 계산의 편의를 위해 로그 가능도(log likelihood)함수로 치환하여 계산한다. 그 이유는 로그 함수는 단조증가(monotone increasing)하기 때문에 가능도에서 극값을 가지는 위치와 로그 가능도에서 극값을 가지는 위치가 같기 때문이다. 따라서 토픽 t가 주어지면, 문단 s의 연관성은

$$s^* = \text{argmax}_s \sum_{w\in t}\log\left(\prod_{w\in t}\frac{g(w)}{\sum_{w'\in t\cap text(s)} g(w')}\right) \quad (12)$$

수식 (12)를 사용하여 토픽과 문단이 얼마나 연관성이 있는지를 계산하게 된다. 이러한 과정을 통해 각 토픽(휴대폰의 주요 특징)은 가장 연관성이 있는 문단들을 매칭 할 수 있게 된다. 그리고 문단이 속해 있는 리뷰 포스트를 찾아서, 특징과 리뷰 포스트를 연결시킨다.

5. 진화 그래프 생성

진화 그래프는 노드와 링크로 구성된다. 각 노드는 특정 휴대폰을 의미하고 두 휴대폰간의 특징 상속(feature inheritance)가 발생한다면, 두 노드간에 링크를 연결한다. 예를 들면, t_0(2009년 4분기)에 출시된 휴대폰 $Phone_{t_0}$과 t_1(2010년 1분기)에 출시된 휴대폰 $Phone_{t_1}$이 있다고 하자. 그리고 $Phone_{t_0}$의 주요특징은 f_1,f_2,f_3,f_4이고, $Phone_{t_1}$의 주요특징은 f_1,f_2,f_3,f_5라고 한다면, $Phone_{t_0}$과 $Phone_{t_1}$는 서로 매우 유사한 특징을 가지고 있다고 할 수 있으며, $Phone_{t_0}$에서 $Phone_{t_1}$로 특징상속이 발생한다고 정의한다. 두 휴대폰간의 주요 특징의 유사 정도를 자카드 유사도(Jaccard similarity)와 쿨백-라이블러 발산(Kullback-Leibler divergence)를 통해 구할 수 있다. 예를 들면, $\text{sim}(Phone_{t_0}, Phone_{t_1}) = \frac{|\{f_1,f_2,f_3,f_4\}\cap\{f_1,f_2,f_3,f_5\}|}{|\{f_1,f_2,f_3,f_4\}\cup\{f_1,f_2,f_3,f_5\}|} = \frac{3}{5} = 0.6 \geq \theta$ 이면, $Phone_{t_0}$과 $Phone_{t_1}$는 특징상속이 발생한다고 말한다. 샘플 데이터를 테스트해본 결과, 자카드 유사도와 쿨백-라이블러 발산 방법의 큰 차이는 발견되지 않아, 자카드 유사도를 사용하여 그래프를 생성하였다.

시기별로 이러한 과정을 반복하여 특정 시기에 출시된 휴대폰을 그래프의 노드로 하고, 유사한 특징을 가진 노드들을 시간 순서대로 연결하여, 그래프를 구성할 수 있다. 완성된 그래프는 시간에 따른 휴대폰 간의 진화패턴 (evolutionary structure)를 보이는 휴대폰의 진화계통도라고 할 수 있다.

IV. 사례 연구

1. 데이터 수집 및 구현

그림 5. 세티즌 웹사이트에 올라온 리뷰 포스트의 예제 화면
Fig. 5. An example of review posts in Cetizen

휴대폰과 스마트폰에 대한 국내의 주요 리뷰 사이트인 세티즌과 오픈모바일에서 2001년부터 2012년까지 게재된 약 50,000개의 리뷰 포스트들을 크롤링(crawling)하여 자동으로 수집하였다. 그림 5는 세티즌의 리뷰 포스트를 캡쳐 한 화면이다. 리뷰 포스트는 제목, 날짜, 본문 텍스트로 구성되어 있다. 국민대의 한글 형태소 분석기를 사용하여, 리뷰 포스트 텍스트를 정제하고 MySQL 데이터베이스에 저장 관리하였다. 그리고 JGibbLDA 오픈 소스[29]를 수정하여 텍스트로부터 토픽을 추출하였다. 또한 Algorithm 3과 진화 그래프를 생성하는 프로그램은 자바 프로그래밍 언어로 직접 구현하였다.

IV.2와 IV.3 장의 결과는 출력된 데이터를 바탕으로 시각 디자이너의 도움을 받아 새로이 디자인하였고, 도메인 전문가들의 도움을 받아 나온 결과를 분석하고 정리하였다.

2. 휴대폰의 진화패턴 분석

제안방안에 의해 만들어진 2011년~2012년의 휴대폰의 진화계통도를 통해 인공물의 진화 원리에 대해 자세히 살펴본다.

그림 9에 의하면 조상관계로 맺어진 휴대폰들이 가지고 있는 핵심적인 공통 특징이 유전되는 것을 쉽게 볼 수 있다. 예를 들면, 팬택의 베가 X와 LG 옵티머스 블랙은 '디자인'이라는 특징으로 서로 연결되는 것을 알 수 있다. 한편 LG 옵티머스 2X와 팬택 베가S에서는 오히려 기능이 퇴화된 형태로 진화하였다. 이처럼 과거의 휴대폰이 더 좋은 경우 반대 진화가 일어날 수 있다. 이것은 제조사들이 더 좋은 스펙의 휴대폰만 출시하지 않고, 가격이 저렴하면서 기본 기능만 갖춘 보급형 휴대폰을 출시하기 때문이다. 이것은 생물체의 진화에서도 쉽게 볼 수 있는 현상이다. 예를 들면, 인간의 후각 능력을 개과 동물과 비교했을 때 크게 퇴화했지만 시각 능력은 더욱 발전하였다. 진화계통도에 따르면 휴대폰의 진화는 주로 사용자의 선호에 의해 발생하지만 진화에 큰 영향을 미치는 큰 축은 제조사인 것으로 판명되었다. 다시 말해 삼성 또는 LG 같은 제조사들은 사용자의 선호도를 파악하여 제품을 출시하기도 하고, LTE와 같은 새로운 기술을 등장시켜 주요 트렌드의 흐름을 바꾸어 놓기도 한다. 이러한 사례는 기술의 점진적인 변화나 발전을 나타내는 변이에 해당한다고 볼 수 있다. 최근 연구에 따르면 기술과 생물의 진화는 모듈성(modularity)가 증가하는 방향으로 진화가 이루어진다는 주장이 부각되고 있다. 생물체는 통합성(integration)과 소포화(parcellation)를 통해 모듈성을 증가시켰고, 인공물에 있어 모듈성은 부속품을 대체하고 재조립할 수 있게 함으로써 더 잘 진화할 수 있게 되었다. 예를 들면, 3G망을 사용하는 휴대폰이 있다고 하자. 동영상을 다운로드 받으려면 많은 시간이 필요하다. 마침 LTE 기술이 나왔다면, 휴대폰에서 3G 모듈을 제거하고 LTE 모듈로 재조립하면 간단하다. 만일 이러한 모듈화가 되어 있지 않으면, 휴대폰을 새로 개발해야 되고 시장에 대응하기에 여러 제약이 있을 것이다. 이러한 모듈성을 바탕으로 이미 존재하는 기술들을 조합하여 참신한 기술(변이)가 발생하게 된다. 또한 변이의 일종인 기술 조합(combinatorial evolution)이 수시로 나타난다. 그림 9에서 팬택의 베가 넘버5는 LG 옵티머스 블랙의 특성인 '배터리 성능 향상'과 삼성의 넥서스S의 '화면 크기' 특성들이 조합하여 진화되고 있음을 볼 수 있다. 한편 자연(시장)선택은 유전자와 환경(시장)의 적합도 함수로 나타낼 수 있으며, 만약 시장에서 어떤 제품의 인기가 높아 지속적으로 살아남는다는 것은 자연에서 환경에 잘 적응하여 유전자가 살아남아 계속 복제되는 것과 유사하

다. 어떤 휴대폰이 인기가 높은 것은 그 휴대폰의 모든 특징들이 대중에게 사랑받을 정도로 완벽한 것이 아니라 몇몇 특징들이 대중의 관심을 사로잡는 것이 된다. 예를 들면, 휴대폰에 디지털 카메라를 장착한 것은 큰 인기를 가져왔는데, 디지털 카메라라는 특징을 만들어내는 유전자(디지털 카메라의 속성과 기능)이 계속해서 대중들에게 사랑을 받는 것이다.

이와 같이 자연선택의 변이, 적합도, 유전 조건이 인공물, 특히 휴대폰의 진화에도 잘 적용된다는 것을 알 수 있다. 반면에 생물 진화와 달리, 인공물에서만 나타나는 특성도 발견할 수 있다. 예를 들면, 인간이 쉽게 받아들이는 기술일수록 환경(시장)에서 더 잘 적응한다. 이것은 제품이 최적화되어 설계되어 완벽한 제품일지라도 대중이 받아들이지 않고 외면한다면, 사라질 수 있다. 또한 연예인의 제품 광고 혹은 지인들이나 많은 사람들이 제품을 선호할 때 막연히 받아들이는 점은 인간의 심리 활동이 제품의 진화에 영향을 끼친 것으로 자연의 진화와는 확연히 다르다. 또한 인공물은 진화가 정지된 안정기가 없으며 살아있는 동안 많이 복제하거나 복제하는 동안 오류가 많지 않아야 되는 생물의 진화 원리를 따르지 않는다. 이를 테면, 카메라라는 특징은 2011년 1월에서 2012년 12월에 이르기까지 점점 진화되는 흐름을 보여주는 것이 아니라 특정 분기에서만 진화된다. 이는 대체로 사용자들이 특정 기능이 어느 정도 만족할 만한 수준에 도달하면, 다른 새로운 기능에 관심을 보이기 때문이다. 따라서 어느 한 특징이 발전하면서 지속되는 경우는 드물다. 그리고 결정적으로 진화에 있어 인공물의 경우 인간의 의도가 개입된다는 점에서 생물의 진화 논리로 설명되지 않는 부분이 존재한다. 더욱이 생물체의 진화계통도와 비교할 때, 다른 휴대폰보다 더 이른 시점에 등장한 휴대폰을 알 수 있고, 여러 자손과 조상(multi-furcating)을 가질 수 있으며, 진화계통도 내에 중간 다리 역할을 하는 인터널 노드(internal node)가 있는 것이 특징이며, 본 연구에서 제안한 방안으로 만들어진 휴대폰의 진화계통도는 이러한 특징을 잘 보여준다.

3. 연도별 휴대폰의 특징 분석

이번 절에서는 휴대폰의 진화계통도를 바탕으로 연도별 휴대폰의 특징 및 트렌드 분석을 통해 국내 휴대폰 시장의 변화를 살펴본다. 휴대폰의 특징을 통해 특정 유전자의 진화를 간접적으로 확인할 수 있으며, 진화 경향을 토대로 차세대 휴대폰의 특징에 대한 예측을 시도해볼 수 있다.

먼저 최근(2008년~2012년)에 출시된 휴대폰에 대한 진화패턴을 분석한다. 2008년(그림 6 참조)에 '풀터치'라는 특징이 처음 등장했다. 사용자들이 터치에 관심을 갖기 시작했고, 햅틱 2(SPH-W5500)와 같이 대표적인 풀터치폰은 사용자들의 선택을 받아 여러 분기에 걸쳐 지속적으로 수요가 발생하였다. 또한 카메라에 대한 사용자의 관심에 대한 예로 캔유파파라치폰(CanU801ex)을 들 수 있다. 캔유파파라치폰은 500만 화소로 카메라 화소에 초점을 맞춘 휴대폰이다. 이 폰은 카메라 화소뿐만 아니라 외장 메모리의 지원용량이 증가하였는데 이는 고화질의 사진을 저장하는 것을 용이하게 하기 위해서이다. 또 하나의 특징은 블랙라벨 시리즈 중 하나인 시크릿폰(LU6000)의 지속적인 수요이다. 이것은 블랙라벨 시리즈를 사용자들이 선호했다는 것을 나타낸다. 블랙라벨이란 소재를 고급화하고 가격을 한 단계 높인 고급 의류 제품을 뜻한다. LG전자에서 이를 휴대폰 이름에 접목시켜 2005년 초콜릿폰(LG-KV5900)을 시작으로 특화된 디자인의 휴대폰을 선보이기 시작했다. 특정 시리즈에 대한 사용자의 선호가 휴대폰의 지속적인 수요에 영향을 미쳤다.

2009년(그림 7 참조)에 가장 큰 특징은 '터치'이다. 2008년부터 이어져서 2009년까지 터치에 대한 사용자의 선호가 높았음을 알 수 있다. 1~2분기 사이에 터치와 관련된 특징이 빈번하게 등장한다. 터치 기능이 추가된 것뿐만 아니라 터치 방식의 변화도 볼 수 있다. 휴대폰 터치방식에는 크게 감압식과 정전식 두 가지가 있다. 감압식은 터치스크린의 압력을 감지해 터치 여부를 판단하는 방식이고, 정전식은 화면 위에 손가락을 가만히 대기만 하면 터치를 인식해낼 수 있는 방식이다. DIVX라는 특징들도 있는데, 별도의 인코딩 작업 없이 휴대폰에서 동영상 재생을 가능하게 하는 것을 말한다. 그리고 또 하나 눈에 띄는 것은 뉴초콜릿폰(LG-LU6300)과 맥스폰(LG-LU9400)이다. 뉴초콜릿폰은 카메라 특징들이 결합된 휴대폰으로 사용자들의 선호에 따라 선택을 받았다고 할 수 있다. 그리고 맥스폰은 다수의 휴대폰들의 디자인적 요소와 카메라 화소 등의 특징들이 모두 결합된 휴대폰이라고 할 수 있다.

2009년부터 이어져온 맥스폰(LG-LU9400)의 선호가 4월 ~ 6월의 휴대폰들에도 많은 영향을 끼치는 것을 알 수 있다. 2010년부터는 AP에 대한 사용자의 관심이 증가하면서 옵티머스 2X(LG-SU660)와 같은 듀얼코어를 탑재한 폰이 선호됨을 알 수 있다. 또 눈에 띄는 특징은 '아몰레드(AMOLED)'로, 기존 LCD와 다르게 백라이트를 필요로 하지 않는 특징 때문에 제품의 두께를 더욱 얇게 만들 수 있는 디스플레이 패널 방식이다. 얇다는 장점뿐 아니라 기존 LCD 방식보다 명암비가 높고 응답속도도 빠르다는 장점을 가지고 있다. 삼성전자에서 아몰레드 시장 점유율이 꽤 높았기 때문에 아몰레드라는 용어를 마케팅으로 사용하기 시작했고, 사용

자들이 휴대폰을 선택할 때 큰 작용을 한 것으로 보인다. 옵티머스Q(LG-LU2300)과 아이폰4의 특징들에서 알 수 있듯이 UI, 키패드와 같은 하드웨어적인 향상도 사용자 선호에 영향을 준다는 것을 알 수 있다.

그림 8은 2010년 휴대폰의 진화계통도를 도식화한 그림이다. 2011년 ~ 2012년도 휴대폰 진화계통도는 3가지 주요 특징들을 보인다. 첫 번째 특징은 디스플레이다. 화면 크기, LCD, 해상도의 특징을 모두 포함한다. 이 중에서도 특히 휴대폰의 화면크기에 대한 사용자들의 선호도가 휴대폰 진화에 영향을 미쳤다는 것을 알 수 있다. 실제로 삼성은 사용자의 선호도의 변화에 발맞추어 2011년도 말 5인치 급의 갤럭시 노트를 출시했고, 그 뒤로 LG 옵티머스뷰 시리즈와 팬택 베가 S5와 같은 휴대폰들이 등장하기 시작했다. 두 번째로 두드러지는 특징은 속도와 관련된 LTE와 AP이다. LTE는 통신속도를 의미하며, AP는 스마트폰, 디지털TV 등에 사용되는 스마트폰의 중앙처리 장치라고 할 수 있다. 점차 3D 게임과 같은 고사양 어플리케이션을 원하는 사용자들이 증가했고, 좋은 성능의 AP가 탑재된 스마트폰에 대한 선호도가 증가하고 있는 것을 알 수 있다. 마지막 특징은 카메라이다. 이 시기에 사용자가 휴대폰을 선택할 때, 카메라의 화소나 기능도 중요한 고려 요소 중 하나인 것을 알 수 있다. 2011년 이전에도 맥스폰, 뉴초콜릿폰 등 500만 화소의 고화질 휴대폰이 등장하긴 했지만, 2011년 ~ 2012년도에 이르러서 고화질 휴대폰이 보급화 되었다는 것을 알 수 있다. 또한 화소 이외에 스마트 오토포커스 기능, 듀얼 LTE 플래시, 손떨림 보정 등 디지털 카메라의 기능이 추가되어 휴대폰의 변천하는 패턴을 엿볼 수 있다.

2001년부터 2007년 사이에 출시된 휴대폰에 대한 진화패턴을 정리하면, 2001년에는 휴대폰의 부가적인 기능이나 감성적인 측면에서 사용자의 선호가 나타났고 슬라이드 형태의 휴대폰이 등장하기 시작했다. 2002년 시기에는 휴대폰의 디스플레이의 변화가 조금씩 일어났다. 2003년에는 듀얼 카메라의 휴대폰이 없었던 시기이기 때문에 앞뒤로 모두 사진을 찍고 싶어 하는 사람들의 수요가 늘어났고, 64화음의 벨 소리를 가진 애듀폰(SCH-E250)이 등장하였다. 2004년에는 카메라 화소에 대한 사용자의 관심이 매우 컸으며, 인테나(내장형 안테나를 사용해 외형적으로 안테나를 보이지 않도록 만든 형태)의 특징들이 나타났다. 2005년에는 '카메라 화소'라는 특징이 두드러지게 나타났으며, T 슬라이드폰(PT-S110)이나 가로본능(SCH-B250)와 같은 로테이션 형태를 선호하기 시작했다. 또한 블랙라벨 시리즈 초콜릿폰은 사용자들의 지속적인 선택을 받았으며, 초슬림 휴대폰의 특징이 나타나기 시작했다. 2006년도 마찬가지로 크레이저(MS700)과 같은 슬림하고 콤팩트한 디자인의 휴대폰이 각광을 받았으며, DBM에 대한 사용자의 관심이 증가하면서 DMB의 유무는 휴대폰을 선택하는 중요한 요소가 되었다. 2007년에는 슬림에 대한 사용자 선호가 계속 증가하였고, 샤인폰(LG-SV420)에 대한 지속적인 수요가 있었다. 전체적으로 슬림한 디자인을 추구하면서, 사용자들이 휴대폰 디자인에 대한 많은 관심을 반영되었다.

V. 결 론

본 논문에서는 휴대폰의 진화패턴 그래프를 생성하는 알고리즘을 제안하였다. 국내 주요 휴대폰 리뷰 사이트에서 사용자가 작성한 리뷰 포스트 데이터를 수집하여, 휴대폰의 특징들을 추출하고, 휴대폰간의 특징들의 유사 정도를 판단하여 휴대폰의 진화계통도를 완성하였다. 이를 바탕으로 휴대폰의 진화패턴을 분석하고, 그 의미를 고찰하였다. 그리고 연도별 특징 및 트렌드 분석을 통해 국내 휴대폰 시장의 변화를 살펴보았다. 연구 결과에 따르면 휴대폰의 진화도 생물체의 진화와 비슷한 양상을 보인다는 사실을 발견하였다. 이를 테면, 휴대폰은 사용자 선택, 제조사에 의한 새로운 기술 등장, 특징 상속 등을 통해 변천되는데, 이것은 생물의 진화논리인 자연선택, 돌연변이, 유전 조건들과 맥을 같이한다고 볼 수 있다. 또한 생물 진화와 달리, 인공물에서만 나타나는 특성도 살펴보았다.

향후, 본 연구는 확장이 필요하다. 우선, 정답 세트가 존재하지 않기 때문에, 제안방안을 직접 평가할 수 없으나, 다른 관점에서 시도된 휴대폰의 진화계통도와 비교 분석을 통해 간접적으로 제안방안을 평가할 수 있을 것이다. 또한 자연어 처리를 향상시키고, 소셜 네트워크 서비스 등 다른 데이터 소스(source)를 추가하여 데이터 커버리지(coverage)를 높인다면, 좀 더 정확한 결과를 얻을 수 있을 것으로 예상한다.

REFERENCES

[1] Arthur, W., "Increasing returns and path dependency in the economy," University of Michigan Press, 1994

[2] Basalla, G., "The evolution of technology," Cambridge University Press, 1988

[3] Best, M., "Models for interacting populations of

memes: Competition and niche behavior," Journal of Memetics, Vol. 1, No. 1, pp.1-9, 1997

〔4〕 Blei, D., Ng, A., Jordan, M., and Lafferty, J., "Latent Dirichlet Allocations," Journal of Machine Learning Research, Vol. 3, No. 4-5, pp.993-1022, 2003

〔5〕 Blei, D. and Lafferty, J., "Dynamic topic models," International Conference on Machine Learning, New York, USA, 2006

〔6〕 Blei, D., "Probabilistic topic models," Communications of the ACM, Vol. 55, No. 4, pp.77-84, 2012

〔7〕 Borshchev, A., and Filippov, A., "From system dynamics and discrete event to practical agent based modeling: Reasons, techniques, and tools," International Conference of System Dynamics Society, July 25-29, Oxford, England

〔8〕 Callon, M. and Law, J., "After the individual in society: Lessons on collectivity from science, technology and society," Canadian Journal of Sociology, Vol. 22, No. 2, pp.165-182, 1997

〔9〕 Cho, H. and Yang, S., "The study on the effects of system quality of smart phone on use of intention," Journal of the Korea Society of Computer and Information, Vol. 16, No. 5, pp. 147-152, 2011

〔10〕 Choi, J., Baek, Y., and Han, S., "A study of receptive factors of smart phones service from the user's perspective," Journal of the Korea Society of Computer and Information, Vol. 18, No. 11, pp. 181-190, 2013

〔11〕 Dalvi, N., Kumar, R., Pang, B., and Tomkins, A., "Matching reviews to objects using a language model," Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, August 2009

〔12〕 Fei-Fie, L. and Perona, P., "A Bayesian hierarchical model for learning natural science categories," IEEE Computer Vision and Pattern Recognition, 2005

〔13〕 Horx, M., "Technolution," 21st Centaury Books, 2009

〔14〕 Hughes, T., "Networks of power: Electrification in western society, 1880-1930," Johns Hopkins University Press, 1983

〔15〕 Kam, M. and Song, M., "A study on differences of contents and tones of arguments among newspapers using text mining analysis," Journal of Intelligent Information Systems, Vol. 18, No.3, pp.53-77, 2012

〔16〕 Kang, S., "Korean lexical analysis," http://nlp.kookmin.ac.kr/HAM/kor/ham-intr.html, 2012

〔17〕 Khanafiah, D. and Situngkir, H., "Visualizing the Phylomemetic tree - Innovation as evolutionary process," Journal of Social Complexity, Vol. 2, No. 2, 2006

〔18〕 Lee, C., Hur, J., Oh, H., Kim, H., Ryu, P., and Kim, H., "Technology trends of issue detection and predictive analysis on social big data," Electronics and Telecommunications Research Institute, 2013

〔19〕 Lee, H., "The evolution of products," Journal of Korean Society of Design Science Vol. 44, 2001, pp.137-146

〔20〕 Loewy, R., "Industrial design," Faber & Faber, 1979

〔21〕 Mei, Q. and Zhai, C., "Discovering evolutionary theme patterns from text - An exploration of temporal text mining," SIGKDD 2005, August 21-24, Chicago, Illinois, USA

〔22〕 Misra, H., Yvon, F., Jose, J., and Cappe, O., "Text segmentation via topic modeling: An analytical study," International Conference on Information and Knowledge Management (CIKM), Hong Kong, China, November 2009

〔23〕 Nelson, R. and Winter, S., "An evolutionary theory of economic change," Belknap Press, 1982

〔24〕 Oh, H., Choi, Y., Shin, W., Jeong, Y., and Myaeng, S., "Trend properties and a ranking method for automatic trend analysis," Journal of KIISE : Software and Applications Vol. 36,

No. 3, 2009, pp.236-243
〔25〕 Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith P., "The author-topic model for authors and documents," Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2004
〔26〕 Sun, B., Mitra, P., Zha, H., and Giles, C., "Topic segmentation with shared topic detection and alignment of multiple documents," SIGIR, Amsterdam, Netherlands, July 2007
〔27〕 Wagner, C., "Topic models," http://www.slideshare.net/clauwa/topic-models-5274169, 2012
〔28〕 Recorded Future, https://www.recordedfuture.com, 2012
〔29〕 JGibbLDA: A Java implementation of Latent Dirichlet Allocation (LDA) using Gibbs sampling for parameter estimation and inference, http://jgibblda.sourceforge.net, 2012

저 자 소 개

온 병 원

1998: 안양대학교
컴퓨터공학과 공학사
2000: 고려대학교
컴퓨터학과 이학석사
2007: 펜실베이니아주립대학교
컴퓨터공학과 공학박사
2008: 브리티시컬럼비아대학교
컴퓨터과학과 포스닥연구원
2010: 일리노이대학교(어바나-샴페인)
차세대디지털과학센터
선임연구원
2011: 차세대융합기술연구원
공공데이터연구센터
선임연구원
2013: 차세대융합기술연구원
공공데이터연구센터 센터장
현 재: 군산대학교
통계컴퓨터과학과 조교수
관심분야: 데이터 마이닝,
데이터베이스,
정보검색, 빅데이터
Email : bwon@kunsan.ac.kr

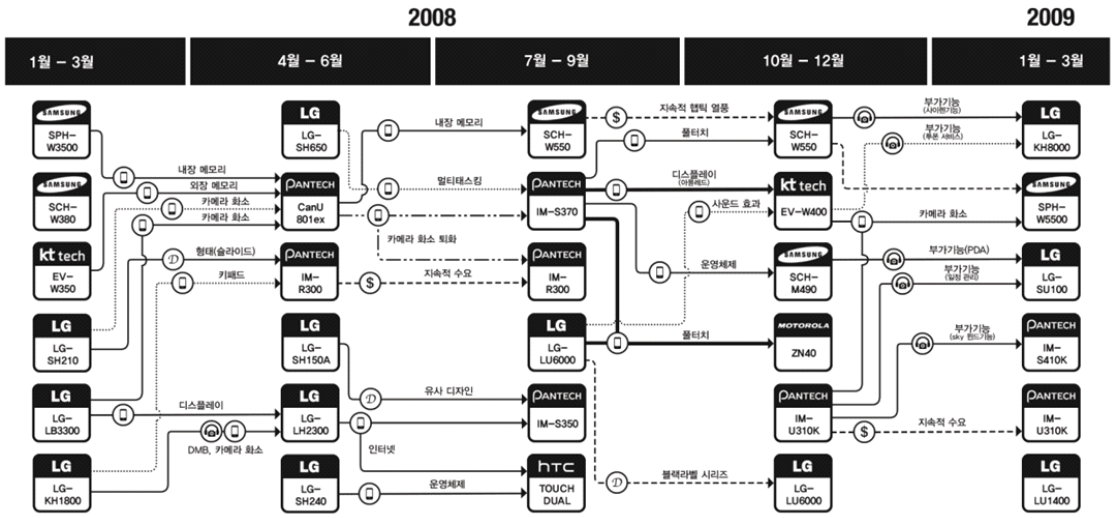


그림 6. 2008년 휴대폰의 진화패턴
Fig. 6. Evolutionary Pattern of Phones in 2008

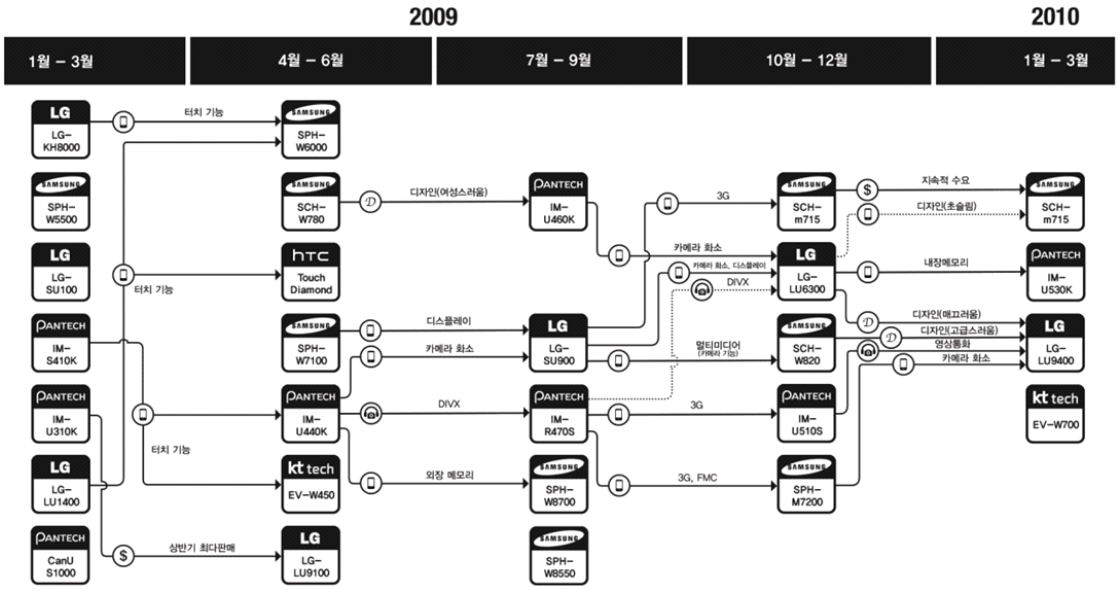


그림 7. 2009년 휴대폰의 진화패턴
Fig. 7. Evolutionary Pattern of Phones in 2009

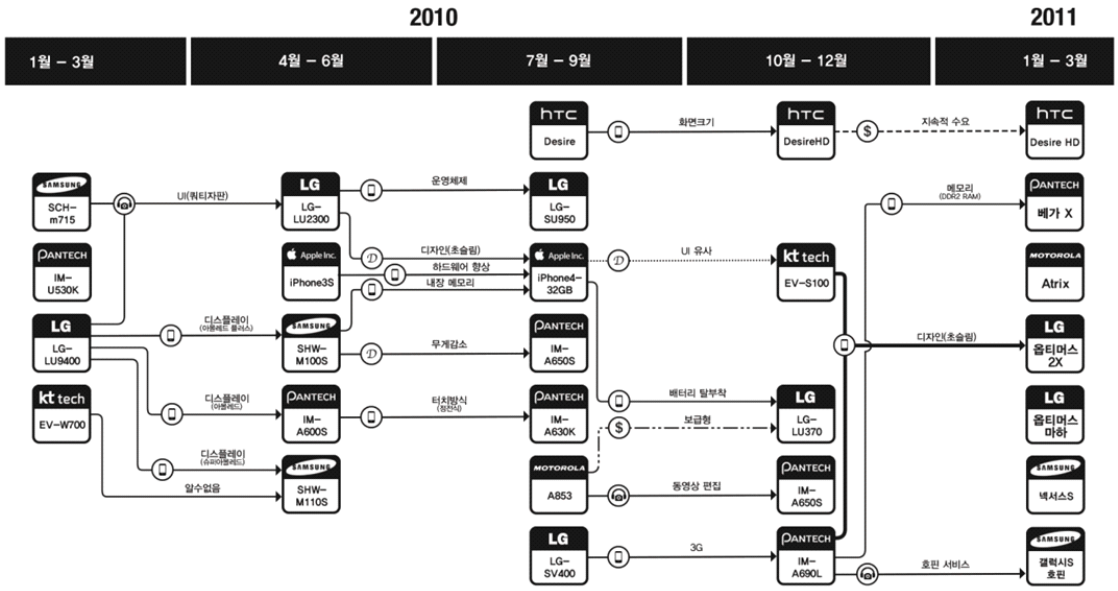


그림 8. 2010년 휴대폰의 진화패턴
Fig. 8. Evolutionary Pattern of Phones in 2010

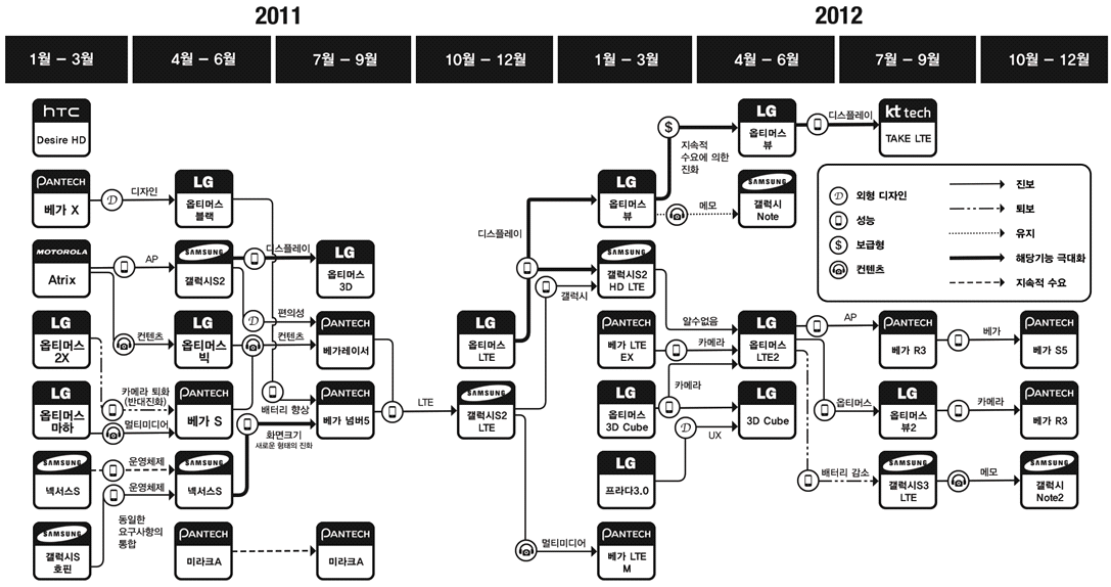


그림 9. 2011~2012년 휴대폰의 진화패턴
Fig. 9. Evolutionary Pattern of Phones between 2011 and 2012