메시지의 상관관계를 이용한 분산병렬처리 기반의 소셜 네트워크 서비스 시각화 방법

김용일* · 박선** · 류갑상***

Visualization Method of Social Networks Service using Message correlations based on Distributed Parallel Processing

Yong-Il Kim* · Sun Park** · Gab-Sang Ryu***

요 약

본 논문은 소셜 네트워크상의 내부관계와 외부관계를 반영하여 사용자간의 관계를 사용자 중심으로 계층적 시각화하는 새로운 클라우드 기반의 방법을 제안한다. 본논문의 시각화방법은 상관관계 행렬을 이용하여 사용자의 내부관계를 계산하여 소셜 네트워크상 사용자 중심의 관계 계층을 잘 나타내며, 소셜 네트워크의 외부 관계를 이용하여 사용자의 계층 관계에 접근 노드의 중요도를 반영한다. 제안방법의 사용자들은 소셜 네트워크상의 사용자 노드 관계가 계층적으로 시각화되기 때문에 사용자 관계를 잘 이해할 수 있다. 이외에 제안된 방법은 하둡(hadoop)과 하이프(hive)를 이용하여 분산저장 및 병렬로 계산하며, 계산 결과는 D3를 이용하여 계층적 그래프로 시각화한다.

ABSTRACT

This paper proposes a new visualization method based on cloud technique which uses internal relationship of user correlation and external relation of social network to visualize user relationship hierarchy. The visualization method of this paper can well represent user-focused relationship hierarchy on social networks by a correlation matrix. The importance of a access node reflects into user relationship hierarchy by exploiting external relation of social network. Users of the method can well understand user relationships on account of representing user relationship hierarchy from social networks. In addition, the method use hadoop and hive for distribution storing and parallel processing which the result of calculation visualizes hierarchy graph using D3.

키워드

소셜 네트워크 서비스, 시각화, 클라우드, 분산병렬처리

Key word

Social Networks Service (SNS), Visualization, Cloud, distributed parallel processing

* 정회원 : 호남대학교
** 정회원 : 국립목포대학교
*** 정회원 : 동신대학교(교신저자, gsryu@dsu.ac.kr)

접수일자 : 2013. 04. 01
심사완료일자 : 2013. 04. 25

Open Access http://dx.doi.org/10.6109/jkiice.2013.17.5.1168

Ⅰ. 서 론

클라우드 컴퓨팅의 정보를 인터넷 상의 서버에 저장하고, 각종 단말기를 통하여 언제 어디서든 서버에 접근하는 것이다. 즉, 구름(cloud)과 같이 무형의 형태로 존재하는 하드웨어 및 소프트웨어 등의 컴퓨팅 자원을 자신이 필요한 만큼 빌려 쓰고 사용요금을 지불하는 컴퓨팅 서비스이다[1].

온라인상에 형성되는 다양한 사회적 네트워크(social networks)의 정보들은 온라인이나 오프라인상의 상업활동의 추천정보나 기타 다양한 분야의 분석정보로 활용될 수 있는 유용한 정보이다. 이 때문에 소셜 네트워크 서비스 분석 및 시각화 방법에 대해서 현재 많은 연구들이 진행되고 있다. 특히 사회관계 중요한 공동체나 중심적인 역할을 수행하는 사용자를 네트워크상에서 그래프로 표현하는 시각화방법이 소셜 네트워크의 중요한 분석방법으로 많은 선호를 받고 있다. 즉, 소셜 네트워크의 시각화 방법을 이용하여서 소셜 네트워크상의 중요한 공동체나 중심적인 역할을 수행하는 사용자의 검색은 다양한 분야의 기초분석 자료로 활용할 수 있다.

현재 소셜 네트워크 시각화를 위한 연구의 대표적인 접근방법으로는 노드 링크(NL, node-link) 접근방법[2], 행렬 그래프(MAT, matrix graph) 접근방법[3], 노드 링크와 행렬 그래프의 혼합형 접근방법(hybrid of NL and MAT)[4, 5]이 주로 연구되고 있으며, 이외에도 다양한 기법들을 기존방법들에 적용하여 성능을 향상시키는 확장형 접근방법들이 있다[6, 7].

이전 연구들을 문제점을 분석해 보면 다음과 같다. 첫째, 복잡한 다차원 그래프를 기반으로 시각화하기 때문에 사용자를 중심으로 한 사회관계의 중요도를 직관적으로 파악하기 힘든 가독성 문제를 가지고 있다. 둘째, 대부분의 시각화 방법들이 네트워크상 노드간의 접근양에 의해서만 사용자 관계를 나타내기 때문에 사용자의 메시지 내용이 상호관계에 반영되는 것이 미흡한 문제를 가지고 있다. 마지막으로, 대분의 시각화 방법들은 시각화에 만 중점을 두고 있기 때문에 기하급수적으로 늘어나는 소셜 네트워크의 빅데이터를 효율적 신속히 처리할 수 없다.

본 논문은 이전 방법들의 문제점을 해결하기 위해서 소셜 네트워크상의 내부관계와 외부관계를 반영하여 사용자간의 관계를 사용자 중심으로 계층적 시각화하는 새로운 클라우드 기반의 방법을 제안한다. 제안방법은 상관행렬로 계산된 내부관계정보와 노드 상호작용 정보에 의한 외부 접근 정보를 이용하여서 사용자중심의 계층적 시각화방법을 제한다. 또한 제안된 방법은 하둡(hadoop)[8]과 하이브(hive)[8]를 이용하여 분산저장 및 병렬로 시각화 처리속도를 향상시켰으며, 계산결과는 D3[9]를 이용하여 계층적 그래프로 시각화한다.

하둡은 분산 파일 시스템(distribution file system)과 분산 컴퓨팅을 위한 맵리듀스(MapReduce)를 포함하여 개발된 분산병렬처리 시스템이다[8]. 하이브는 페이스북이 개발한 하둡 기반의 데이터웨어하우징 시스템으로 SQL과 매우 유사한 HiveQL이라는 쿼리를 제공한다[8]. D3는 마이크 보스탁이 만든 자바스크립트 라이브러리로 데이터 집합의 문맥 안에 HTML(hyper text markup language), SVG(scalable vector graphics), Canvas 같은 웹 페이지 요소를 데이터에 따라 보여주고, 삭제하며, 편집할 수 있도록 지원한다[9].

본 논문의 구성은 다음과 같다. 2장에서는 상관관계와 소셜 네트워크 노드 관계를 이용한 클라우드 기반의 시각화 방법에 대하여 알아보며, 3장에서는 실험 및 분석결과를 설명한다. 마지막 4장에서는 결론에 대하여 기술한다.

Ⅱ. 제안방법

본 논문에서 제안한 클라우드 기반의 분산병렬 시각화 시스템은 그림1과 같이 시각화 방법 모듈, 분산병렬처리 모듈, 시각화 표현 모듈로 구성된다. 시각화의 클라우드 기반의 병렬처리 과정은 사용자가 소셜 네트워크의 시각화 분석을 원하면 시각화 방법 모듈의 그림1(a) 전처리 단계에서 소셜 네트워크의 XML 자료를 미리 정의된 HiveQL 스키마에 적합한 관계형 자료로 변환한다.

변환된 자료는 분산병렬 처리 모듈의 그림1(c) 하이브를 통하여 그림1(d) 하둡에 분산 데이터로 저장되며, 시각화 방법 모듈의 그림1(b) 시각화 알고리즘을 이용하여 하이브를 통해 저장된 자료를 하둡에서 분산병렬로 계산한다. 계산결과를 시각화 표현 모듈인 그림1(e) D3

로 보내어 소셜 네트워크상의 사용자를 계층으로 시각화 표현한다.

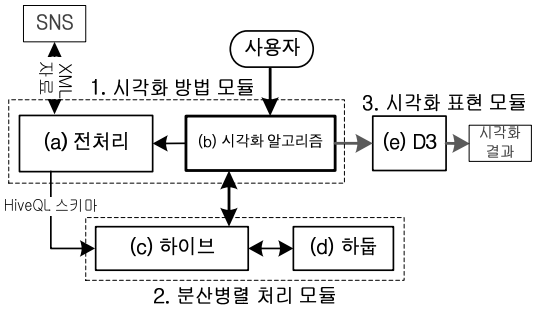

그림 1. 클라우드 기반 시각화 방법
Fig. 1 visualization method based on cloud

2.1. 시각화 방법 모듈

2.1.1. 전처리

그림1(a)의 전처리 단계에서는 소셜 네트워크의 XML 자료를 하이브 스키마에 적합한 관계형 자료로 변환하여 하이브를 통하여 하둡에 분산 데이터로 저장한다. 그림2는 연구 그룹에서 하루 동안에 교류되어 온 이메일 자료[10]를 노드링크로 시각화하여 표현한 예이고, 그림3은 이메일 교류의 XML 자료의 일부를 나타낸 것이다.

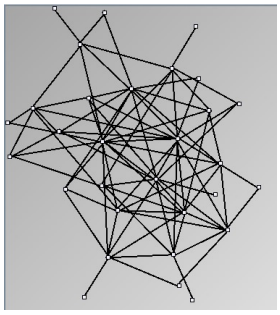

그림 2. 이메일 교류의 노드링크 시각화[10]
Fig. 2 node-link visualization of email exchange[10]

일반적으로 시각화된 소셜 네트워크는 그림3과 같이 XML자료로 저장할 수 있다. 이 때문에 XML자료를 하이브에 저장할 수 있도록 하이브 스키마에 적합한 관계형 자료로 변환한다.

```
<?xml version="1.0" standalone="true"?>
<!DOCTYPE graphml SYSTEM "http://graphml.graphdrawing.org/dtds/1.0rc/graphml.dtd">
- <graphml>
  - <graph edgedefault="directed">
      <key for="node" id="labels">string</key>
      <key for="edge" id="weight">double</key>
    - <node id="n0">
        <data key="labels">ARC SRC For Ultra Broadband Information Networks (CUBIN)</data>
      </node>
    - <node id="n1">
        <data key="labels">ASSET</data>
      </node>
    - <node id="n2">
        <data key="labels">Autonomous Systems & Sensing Technologies</data>
      </node>
    - <node id="n3">
        <data key="labels">Board</data>
      </node>
    - <edge id="e6" target="n12" source="n28">
        <data key="weight">1422.0</data>
      </edge>
    - <edge id="e7" target="n20" source="n26">
        <data key="weight">29.0</data>
      </edge>
    - <edge id="e978" target="n17" source="n9">
        <data key="weight">9.0</data>
      </edge>
    - <edge id="e979" target="n18" source="n9">
        <data key="weight">1.0</data>
      </edge>
    </graph>
</graphml>
```

그림 3. 그림2의 XML 자료[10]
Fig. 3 XML data of figure 2[10]

2.1.2. 시각화 알고리즘

그림1(b)의 시각화 알고리즘은 본 논문의 저자들이 이전에 제안한 시각화 방법[7]을 다음과 같이 확장하였다. 즉, 그림1(2)의 분산병렬 처리 모듈을 통하여 분산병렬로 계산한 후에 계산 결과를 그림1(3)의 시각화 표현 모듈에 전달한다.

저자들의 이전 제안방법은 네트워크상의 내부관계와 외부관계를 반영하여 사용자간의 관계를 사용자 중심으로 계층적 시각화하는 방법이다[7]. 사용자의 내부 상관관계 계산은 소셜 네트워크에서 전송되는 사용자의 메시지 내부의 정보를 얼마나 반영되는지를 나타내는 것으로 식(1)의 상관관계 행렬을 이용하여 계산한다.

$$c_{a,b} = \sum_{t_j} tf_{aj} \times tf_{bj} \quad (1)$$

여기서 t_j는 j번째 열에 속하는 용어들을 나타내며, tf_{aj}는 a번째 행의 메시지에 포함된 j번째 열의 용어들의 출현 빈도를 나타내며, tf_{bj}는 b번째 행의 메시지에 포함된 j번째 열의 용어들의 출현 빈도를 나타낸다.

사용자 외부관계 계산은 소셜 네트워크상에서 참조되는 사용자의 메시지의 양이 사용자들 간에 네트워크상에서 얼마나 반영되었는가를 나타내며 식(2)를 이용하여 계산한다.

$$er(a \rightarrow b) = \left(\frac{nm \times nt}{tm}\right) \times \sum_{i=1}^{td} \left(\frac{d_i}{td-(1-i)}\right) + \left(\frac{nrm}{tm \times ru}\right) \quad (2)$$

여기서 *nm*은 사용자 *a*가 사용자 *b*에게 보내는 메시지 개수, *nt*는 두 사용자가 참조한 모든 메시지에 포함된 명사 용어의 개수, *tm*은 소셜 네트워크상에서 모든 사용자가 참조하는 메시지의 개수이다. *td*는 메시지가 참조되는 총일자의 개수이며, 일자별 참조되는 메시지의 개수 *d*는 최근 날짜를 기준으로 내림차순으로 정렬한다. *nrm*은 사용자가 재전송하는 메시지의 개수, *ru*는 실제 메시지를 재 참조하는 사용자의 개수이다.

사용자의 내부관계와 외부노드관계를 시각화에 반영하기 위해서 내외부관계를 식(3)의 정규화를 이용하여 식(4)와 같이 합산해야 한다.

$$nor(a \rightarrow b) = \frac{u_{ab}}{\sum_{a=1}^{l}\sum_{b=1}^{l} u_{ab}} \quad (3)$$

여기서 *nor*()는 관계의 원소 값을 정규화 시키는 함수며, $a \rightarrow b$는 *a*사용자를 참조하는 *b*사용자를 나타내며, l은 사용자의 총인원수, u_{ab}는 *a*사용자를 참조하는 *b*사용자관계의 원소 값을 나타낸다.

내외부관계의 원소 값을 정규화 한 후에 이들을 시각화에 반영하기 위해서는 내부관계와 외부관계의 원소 값을 다음 식(4)와 같이 합산해야 한다.

$$sie(a \rightarrow b) = nor(c_{a,b}) + nor(er(a \rightarrow b)) \quad (4)$$

여기서 *sie*()는 내부관계와 외부관계 원소를 합산을 시키는 함수이며, *nor*()는 정규화 함수, $c_{a,b}$는 *a*와 *b*사용자에 대한 상관관계 원소, *er*($a \rightarrow b$)는 *a*와 *b*사용자에 대한 외부관계의 원소를 나타낸다.

그림4는 소셜 네트워크의 XML자료를 분산병렬처리를 위한 하이브 스키마에 적합한 관계형 구조로 저장하기 위한 UML(unified modeling language)로 나타낸 것으로 내부관계 및 외부관계 정보에 이용되는 변수 및 함수를 정의하였다.

그림4의 내부 상관행렬계산 함수인 *correlation*()을 HiveQL 문으로 변환하여 분산병렬로 계산한다. 외부관계 역시 그림4의 외부 접근정보 함수 *er*()을 HiveQL 문으로 변환하여 분산병렬로 계산한다. 계산된 내부관계와 외부관계에 정규화 함수인 *nor*()를 HiveQL 문으로 변환하여 정규화하고, 정규화된 내외부관계를 합산하기 위하여 합산함수 *sie*()를 HiveQL 문으로 변환한 후 합산하여 시각화에 반영한다.

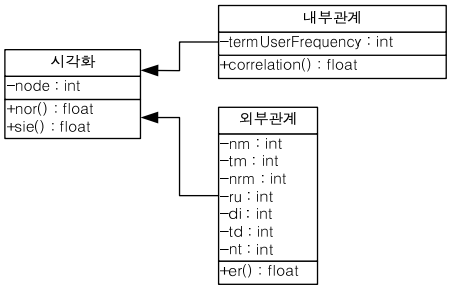

그림 4. 내외부관계의 UML다이어 그램
Fig. 4 UML diagram of internal and external relationship

2.2. 분산병렬 처리 모듈

그림1(2)의 분산병렬 처리 모듈은 그림1(c) 하이브와 그림1(d)의 하둡으로 구성된다. 분산병렬 처리 모듈은 그림1(1)의 시각화 방법모듈의 전처리에서 변환된 자료를 그림1(c) 하이브를 통하여 그림1(d) 하둡에 분산데이터로 저장하며, 그림1(b) 시각화 알고리즘을 이용하여 하이브를 통해 저장된 자료를 하둡에서 분산병렬로 계산한다. 그림1(d) 하둡의 분산병렬 처리 구성환경은 다음 같다.

본 논문의 분산병렬 처리 시스템은 인텔 3i 기반의 4대의 개인용 컴퓨터를 이용하여 구성하였다. 하둡의 분산 서버 구성정보로 4대의 개인용 컴퓨터를 이용하여 네임노드 1대, 보조 네임노드와 데이터노드 공용 1대, 데이터노드 2대로 구성하였다.

2.3. 시각화 표현 모듈

그림1(3)의 시각화 표현 모듈은 시각화 알고리즘의 결과인 사용자의 계층 구조를 JSON(javascript object notation)형태로 변환하고, 이를 D3의 자바스크립트 라이브러리를 이용하여 웹브라우저에서 그래픽으로 시각화하여 표현한다. 다음 그림5는 시각화 표현 모듈에서 그림2의 노드링크로 시각화된 소셜 네트워크를 제안방법으로 시각화한 결과를 보여준다.

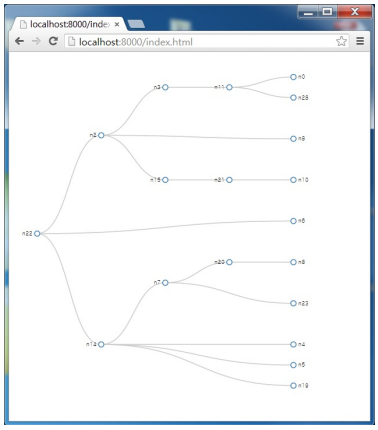

그림 5. 제안방법을 이용한 그림2의 시각화 결과
Fig. 5 Visualization result of Figure 2 using the proposed method

Ⅲ. 실험 및 분석

본 논문에서는 제안 방법의 성능을 평가하기 위하여 Social Network Generation 사이트의 실제 소셜 네트워크 자료인 8개의 각기 다른 규모의 이메일 소셜 네트워크 자료를 이용한다[10]. 본 논문의 평가는 크게 두 가지의 평가척도를 설정하여 평가하였다. 첫 번째 평가척도로 Henry의 표1과 같은 5가지 작업에 대한 평가척도를 이용하여 제시된 시각화 방법을 평가한다. Henry의 평가척도를 기준으로 평가자가 시각화 결과를 1~3 사이의 점수로 평가한다[4]. 두 번째 평가척도는 시각화 방법의 수행시간으로 시각화 결과의 시간을 기록하여 평가한다.

표 1. Henry의 가독성 평가척도
Table. 1 Henry readability measure

평가 작업	내 용
task 1. commonNeighbor	· 두 개의 역할이 주어질 때, 둘 중 직접 연결이 가능한 역할 검색
task 2. shortestPath	· 두 개의 역할이 주어질 때, 최단 경로 검색
task 3. mostConnected	· 가장 관계가 높은 중심 역할검색
task 4. articulationPoint	· 두 개의 부 그래프에서, 절단 점 검색
task 5. largestClique	· 모든 연결들 중에서, 가장 큰 집합의 역할 검색

평가비교는 제안방법인 RH(관계계층; relationship hierarchy)을 NL, MAT, MatLink, MatTrix 방법을 비교하여 평가한다. 여기서 NL은 Ghoniem의 노드링크에 의한 시각화 방법이며[2], MAT는 매트릭스에 의한 시각화방법이고[3], MatLink[4]와 Matrix[5]는 Henry가 제안한 노드링크와 매트릭스의 혼합형 방법이다.

실험1) 첫 번째 실험결과 제안방법인 RH가 NL방법에 비하여 22.1%의 가독성이 우수, MAT방법에 비하여 27.2% 우수, MatLink방법에 비하여 18.6% 우수, MatTrix에 비하여 13.6%가 더 우수하다. 결과를 분석해보면 MatLink와 MatTrix 방법의 시각화 방법이 제안방법과 10%대의 차이를 보이는데 비하여 NL과 MAT 방식은 20% 이상의 차이를 보이고 있다. 이것은 NL방식의 경우 노드와 경로가 증가할 수 록 서로 중복되어 가독성이 떨어지며, MAT방식의 경우 경로가 없기 때문에 평가자들이 판독하기 어려운 것으로 분석된다.

실험2) 두 번째 평가는 8개의 소셜 네트워크에 대한 시각화 방법의 수행시간으로 단독으로 개인 컴퓨터에서 시각화 방법을 수행한 시간과 4대의 개인 컴퓨터에 분산병렬처리하여 수행한 시간을 비교하여 평가한다. 수행 범위는 전처리 시간을 제외한 시각화 알고리즘의 결과가 JSON 파일이 만들어질 때까지이다. 실험결과 단독수행시 62분 30초이며 분산병렬 수행시 21분 9초가 소요되었다. 속도 측정 면에서 제안된 분산병렬처리 방식이 단독방식 보다 평균 41분 21초의 시간 우위에 있다.

Ⅳ. 결 론

본 논문은 이전 소셜 네트워크의 시각화시의 문제를 해결하기 위해서 네트워크상의 내외부관계를 반영하여 사용자간의 관계를 사용자 중심으로 계층적 시각화하는 새로운 클라우드 기반의 방법을 제안하였다. 제안방법은 사용자 중심으로 사용자관계를 계층적으로 표현하기 때문에 소셜 네트워크상에서 중요한 공동체나 중심적인 역할을 수행하는 사용자를 쉽게 찾을 수 있으며, 사용자 관계의 계층적 시각화로 이전 연구들의 가독성 문제를 해결할 수 있다.

또한 제안 방법은 하둡(hadoop)과 하이브(hive)를 이용하여 분산저장 및 병렬로 계산되어 결과는 D3를 이용하여 계층적 그래프로 시각화함으로써 기존의 시각화 방법에 비하여 빠른 결과를 얻을 수 있다.

참고문헌

[1] 클라우드 컴퓨팅, *NAVER 지식백과*, http://terms.naver.com/entry.nhn?cid=200000000&docId=1350825&mobile&categoryId=200000756, 2013, 3.

[2] M. Ghoniem, J. D. Fekete, P. Castagliola, "On the readability of graphs using node-link and matrix based representations, a controlled experiment and statistical analysis, *Information Visualization,* vol. 4, no. 2, pp.114-143. 2005.

[3] S. Wasserman, K. Faust, "Social Network Analysis", *Cambridge University Press,* Combridge, 1994.

[4] N. Henry, J.-D. Fekete, "MatLink: Enhanced Matrix Visualization for Analyzing Social Networks", *LNCS 4663,* Part II, pp. 288-302, 2007.

[5] N. Henry, J.-D. Fekete, M. J. McGuffin, "NodeTrix: a Hybrid Visualization of Social Networks", *IEEE Transactions on Visualization and Computer Graphics,* vol. 13 Issue:6, pp.1302-1309, 2007.

[6] J. Heer, D. Boyd, "Vizster: Visualizing Online Social Networks", *IEEE Symposium on Information Visualziation 2005,* pp.32-39, 2005

[7] 박선, 정종근, 여무송, 이성로, "소셜 네트워크 서비스 사용자의 계층 시각화 방법", 한국정보통신학회 논문지 제16권 제8호, pp.1717-1724, 2012.

[8] 정재화, "시작하세요! 하둡 프로그래밍: 기초부터 실무까지 하둡의 모든 것", 위키북스, 2012.

[9] *D3* (Data-Driven Documents), http://d3js.org, 2013.

[10] *Social Network Generation,* http://www.infovis-wiki.net/index.php/Social_Network_Generation#Real_Social_Networks, 2013.

저자소개

김용일(Yong-Il Kim)

1984년 : 전남대학교(이학사)
1986년 : 한국과학기술원(공학석사)
1986년～1994년 : 한국원자력연구소 선임연구원
1994년～2000년 : 초당대학교 컴퓨터학과 조교수
2002년～현재 : 호남대학교 인터넷콘텐츠학과 조교수
※관심분야 : 빅데이터 처리, 지능형정보검색, 클라우드 컴퓨팅, 지능형 에이전트 등

박선(Park Sun)

1996년 전주대학교(학사)
2001년 한남대학교(석사)
2007년 인하대학교(박사)
2008년～2009년 호남대학교 컴퓨터공학과 전임강사
2010년 전북대학교 인력양성사업단 박사후과정
2010년 12월～현재 목포대학교 정보산업연구소 전임연구교수
※관심분야 : 정보검색, 데이터마이닝, 데이터베이스, 해양IT정보융합

류갑상(Gab-Sang Ryu)

1983년 전남대학교(이학사)
1985년 전남대학교(이학석사)
2006년 고려대학교(이학박사)
1985년～1996년 한국기계연구원 선임연구원
2010년～현재 한국정보통신기술사회 부회장
1996년～현재 동신대학교 컴퓨터학과 교수
※관심분야 : 정보보안, 생산정보화, 무선통신, 융합보안