Service Creation and Research Prediction Based on Knowledge Extraction

Aviv Segev

Department of Knowledge Service Engineering

KAIST

Knowledge Service Engineering

- Too much knowledge
- What can we do with it?
- Create more knowledge
- Extract knowledge to create services
- Mathematical statistics to make general predictions about future behavior

Implementations
Service Creation

Knowledge Maps for e-Learning

Patent Extraction and Trend Prediction

Research Community Prediction in Citation Networks

Knowledge Maps for e-Learning

Jae-Hwa Lee and Aviv Segev, Knowledge Maps for e-Learning, Computers & Education, 59(2), pp. 353-364, 2012

Problem

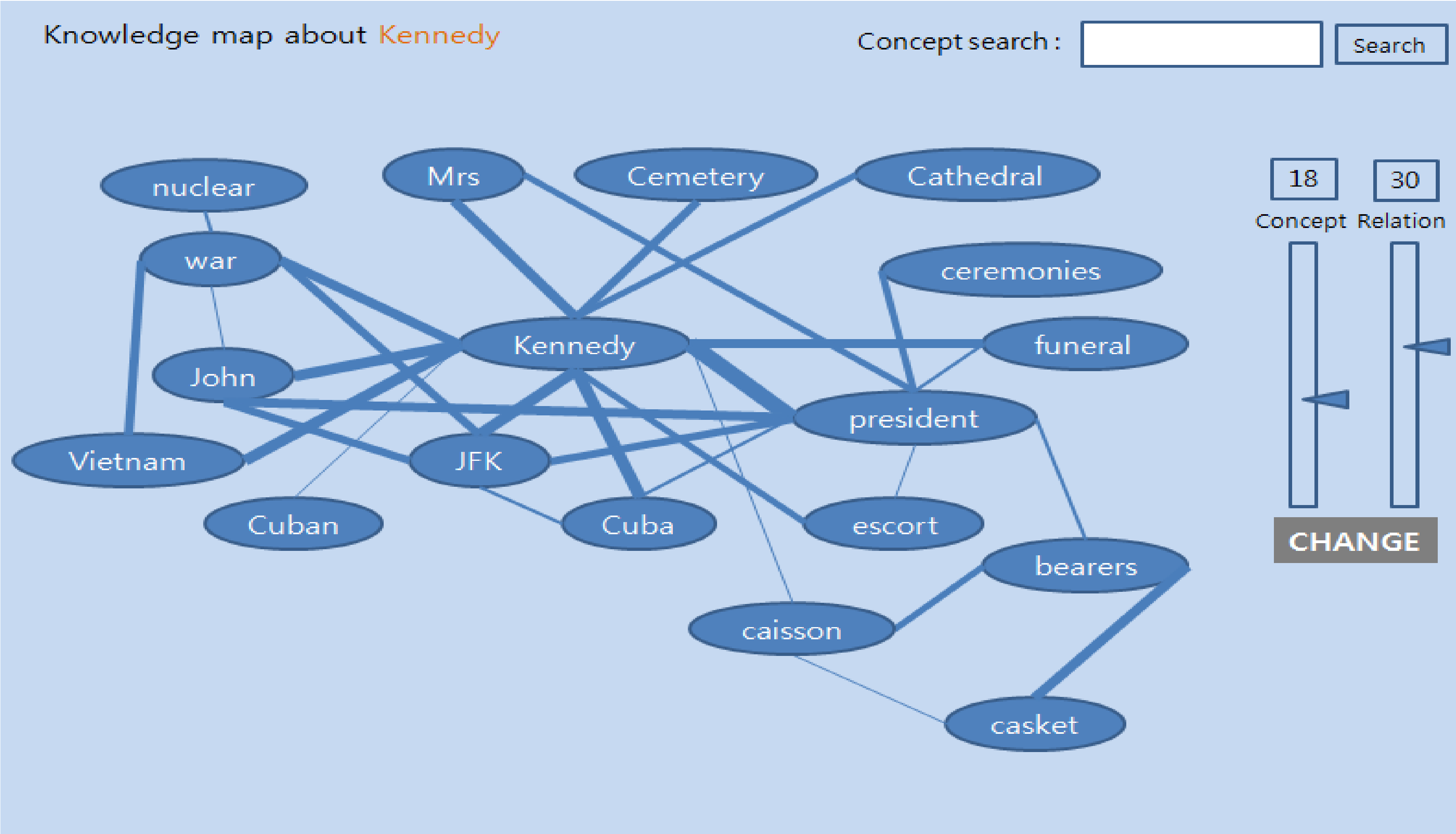
- Learning from text - usually follows the order set by the author, as with reading books

- Create Knowledge Map
- Domain experts are needed.
- Cost of map creation is high.

Solution

- Develop a model which automatically builds a domain knowledge map (K-map) from a set of documents about a specific topic using text mining techniques.

Knowledge Map
John F. Kennedy



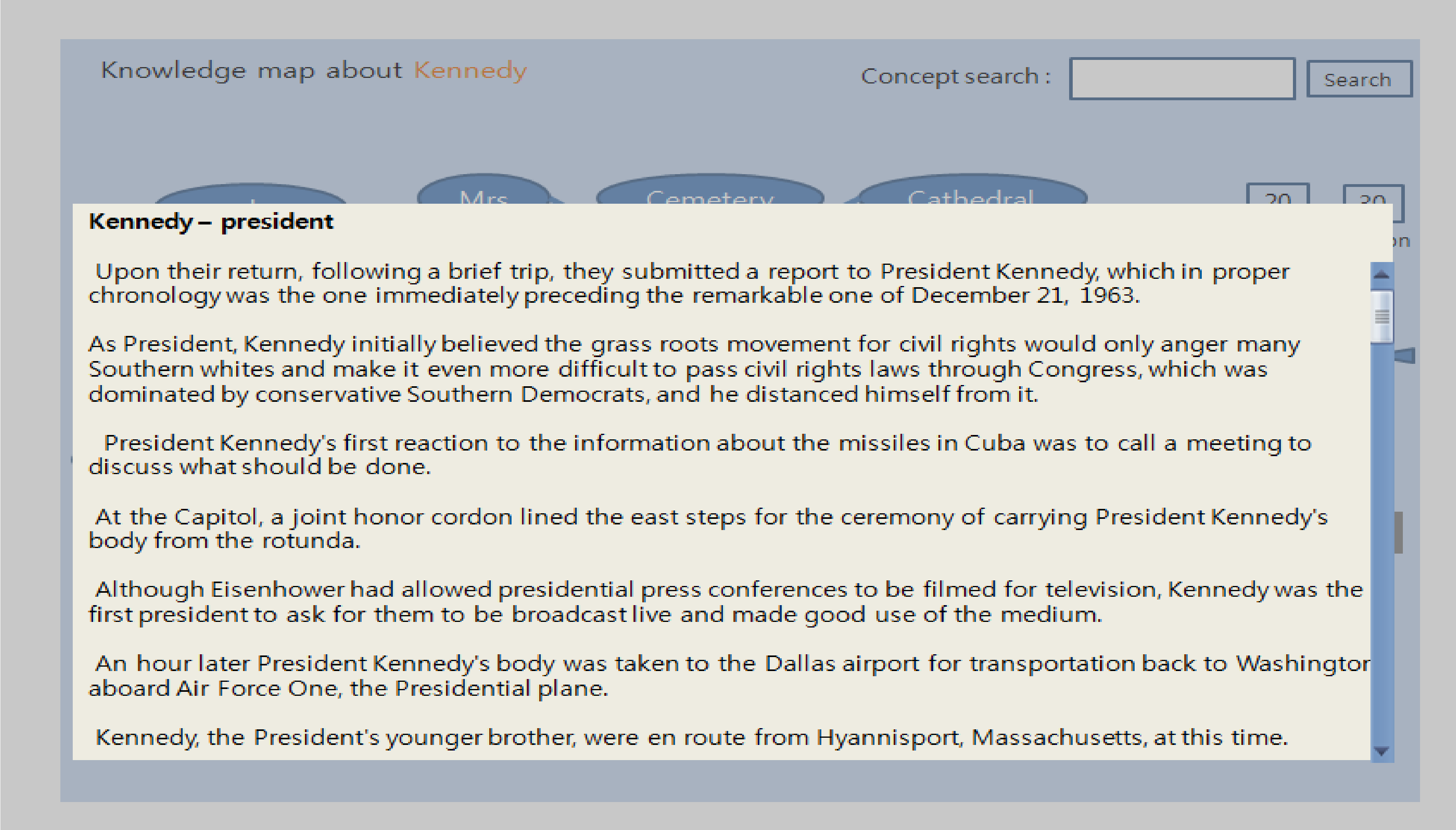
Knowledge map about Kennedy

Concept search :

18 Concept

30 Relation

Sentences Containing ‘Kennedy’ and ‘President’ in K-map Tools



Kennedy – president

Upon their return, following a brief trip, they submitted a report to President Kennedy, which in proper chronology was the one immediately preceding the remarkable one of December 21, 1963.

As President, Kennedy initially believed the grass roots movement for civil rights would only anger many Southern whites and make it even more difficult to pass civil rights laws through Congress, which was dominated by conservative Southern Democrats, and he distanced himself from it.

President Kennedy's first reaction to the information about the missiles in Cuba was to call a meeting to discuss what should be done.

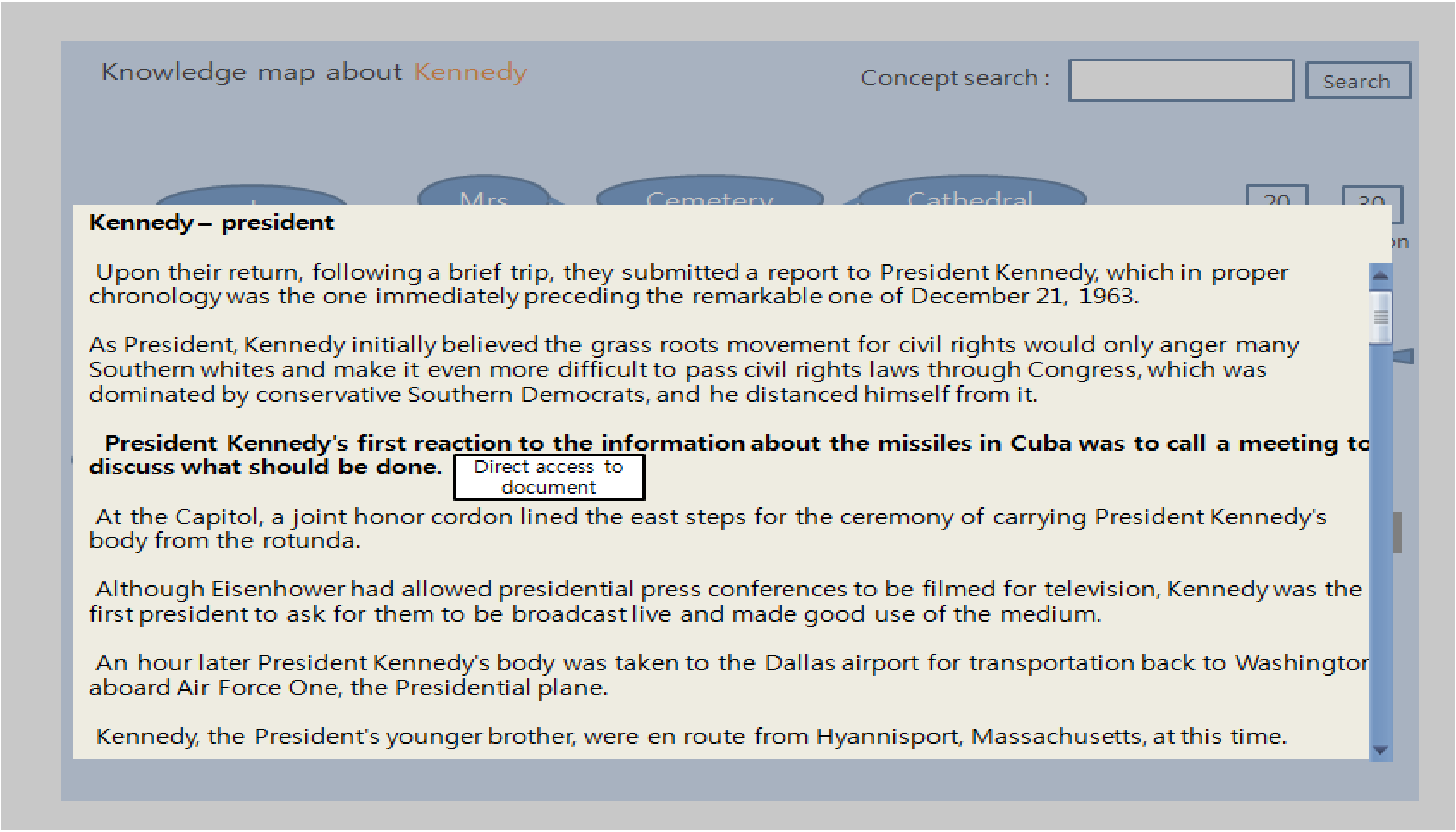
At the Capitol, a joint honor cordon lined the east steps for the ceremony of carrying President Kennedy's body from the rotunda.

Although Eisenhower had allowed presidential press conferences to be filmed for television, Kennedy was the first president to ask for them to be broadcast live and made good use of the medium.

An hour later President Kennedy's body was taken to the Dallas airport for transportation back to Washingtor aboard Air Force One, the Presidential plane.

Kennedy, the President's younger brother, were en route from Hyannisport, Massachusetts, at this time.

Sentences Containing ‘Kennedy’ and ‘President’ in K-map Tools



Kennedy – president

Upon their return, following a brief trip, they submitted a report to President Kennedy, which in proper chronology was the one immediately preceding the remarkable one of December 21, 1963.

As President, Kennedy initially believed the grass roots movement for civil rights would only anger many Southern whites and make it even more difficult to pass civil rights laws through Congress, which was dominated by conservative Southern Democrats, and he distanced himself from it.

President Kennedy's first reaction to the information about the missiles in Cuba was to call a meeting tc discuss what should be done.

At the Capitol, a joint honor cordon lined the east steps for the ceremony of carrying President Kennedy's body from the rotunda.

Although Eisenhower had allowed presidential press conferences to be filmed for television, Kennedy was the first president to ask for them to be broadcast live and made good use of the medium.

An hour later President Kennedy's body was taken to the Dallas airport for transportation back to Washingtor aboard Air Force One, the Presidential plane.

Kennedy, the President's younger brother, were en route from Hyannisport, Massachusetts, at this time.

K-map Benefits

- A user can see key concepts in a domain as well as strongly related concepts.
- As a user reads, he can directly access a document he wants from a certain sentence; in other words, K-map can function as a search engine.
- By exploring the map, a user can learn about the domain at some level of knowledge without accessing original documents. As a user explores a domain K-map, he can see the holistic/overall picture.
- Since a user can choose relations based on keywords, he can selectively learn about the domain, which is hardly possible when learning from text.

Keyword Extraction

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N / n_k)]^2}}$$

W_{ik} : weight of term k in document i

tf_{ik} : term frequency of term k in document i

N : total number of documents

n_k : number of documents that contain term k

$$\mathbf{W_{MT} = Max(W_{D_i T})}$$

D_i = ith document, i = 0,1,2…total number of the documents in K-map

W_T= Weight of term T in K-map

W_{DiT}= Weight of term T in Di

Relation Extraction

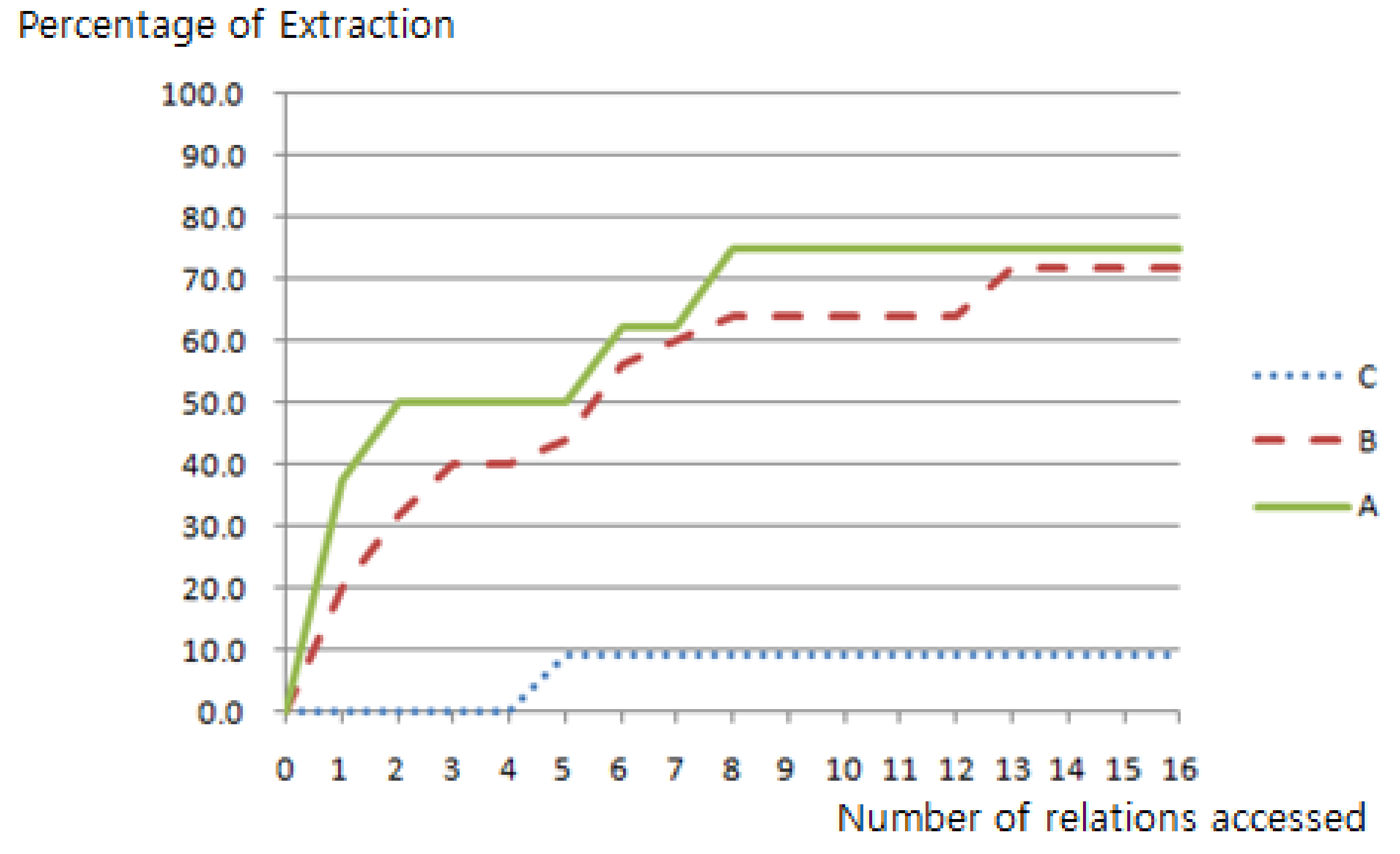
$$\mathbf{R}_{\mathbf{i,j}} = \sum_{\mathbf{D_m}} \sum_{\mathbf{S_n}} \frac{\mathbf{2}}{\mathbf{N}_{\mathbf{D_m S_n}}}$$

i, j = keyword pair
$R_{i,j}$ = score of relation between word i and word j
m = 1,2,… ,Total number of documents in a map
n = 1,2,…, Total number of sentences in document D_m
S_n = nth sentence
D_m = mth document
$N_{D_m S_n}$ = total number of words in sentence S_n, document D_m

Experiments
Categorizing Sentences

- A : Sentences that have main ideas or play a big role in understanding the topic
- B : Sentences that support main ideas or partly help understand the topic
- C : Sentences that are not related to the topic or are not helpful.

Percentage of Sentences Extracted in 3% K-map



Free Recall Experiment

- Participants of the experiment were asked to write everything they learnt from the material after 8 minutes of learning time.
- A grader who does not have any information about groups determined the free recall scores for all the participants.

Comparison of the Amount of Idea Units

Group	Observed Mean	SD	Sample Size
Document Group	15.82	5.29	17
Map Group	15.59	5.64	17
Total	15.71	5.39	34

Amount of Irrelevant Information

Group	Observed Mean	SD	Sample Size
Document Group	3.24	2.44	17
Map Group	0.82	1.29	17
Total	2.03	2.28	34

Results

- K-Map successfully filters out the sentences considered not important to the main idea.
- The results show that there was no statistical difference between the groups recall of important sentences.
- The results showed that with K-map users learned information that is more important, in comparison to the information learned with documents.

Multilingual Knowledge Extraction in Patents

FLOCK – Fuzzy Logic Ontology Context Knowledge

Problem

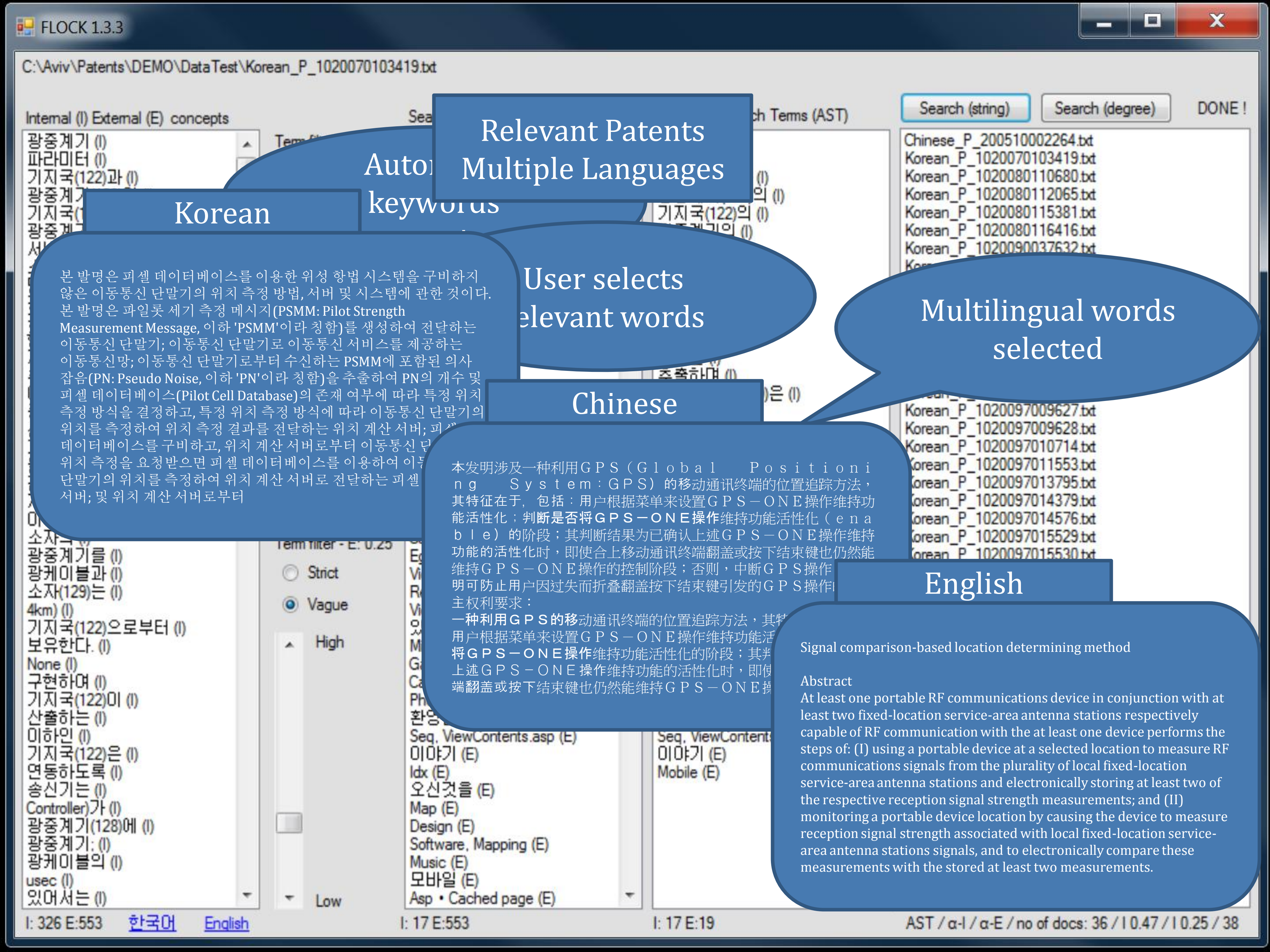
- ➢Decrease the decision time for patent processing (currently 3-4 years)
- ➢Limited to the languages and terms the patent officer knows

Solution

- ➢Semi-automatic knowledge extraction for patent decision makers
- ➢Extract relevant patents in **multiple languages** without the need to translate

Input – Korean Patent

본 발명은 위치 기반 서비스를 제공하는 시스템에 관한 것이다. 본 발명은 위치 기반 서비스를 제공하는 시스템에 있어서, 이동통신 단말기와 통신하여 무선 호의 처리를 위한 제반 기능을 수행하는 무선 기지국; 상기 무선 기지국으로부터 상기 이동통신 단말기로 전송되는 상기 무선 호를 지연하여 중계하는 광중계기; 및 상기 이동통신 단말기로부터 위치 기반 서비스 요청 신호를 수신한 경우, 상기 이동통신 단말기로부터 상기 무선 기지국 또는 상기 광중계기를 매개로 무선환경 파라미터 신호를 수신하고, 상기 무선환경 파라미터 신호로부터 기지국 신호 지연 값 또는 광중계기 신호 지연 값을 추출하며 상기 기지국 신호 지연 값 또는 상기 광중계기 신호 지연 값을 이용하여 상기 이동통신 단말기의 위치 정보를 산출하는 위치 기반 서비스 제공 서버를 포함하는 것을 특징으로 하는 위치 기반 서비스를 제공하는 시스템을 제공한다. 본 발명에 의하면, 통신 사업자 입장에서는 위치 기반 서비스의 측위 정확도를 향상하기 위해 별도의 망 투자 비용을 절감할 수 있는 효과가 있으며, 트래픽의 증가 없이 위치 기반 서비스를 제공할 수 있는 효과가 있다. 최근 공간을 초월하여 인터넷 등의 통신 서비스를 제공하기 위하여 수많은 기업들이 무선 인터넷이라는 새로운 기술 개발에 박차를 가하고 있다. 무선 인터넷은 사용자가 이동하는 중 무선망(Wireless Network)을 통해 인터넷 서비스를 이용할 수 있는 환경과 기술을 말한다. 휴대폰 관련 기술의 발달과 휴대폰 보급률의 비약적인 증가는 이러한 무선 인터넷 환경의 발전을 더욱 촉진시켰다. 한편, 휴대폰이나 피디에이(PDA) 등과 같은 이동통신 단말기를 이용한 다양한 무선 인터넷 서비스들 중 특히, 위치 기반 서비스(LBS: Location Based Service)는 넓은 활용성 및 편리함으로 인하여 크게 각광받고 있다. 위치 기반 무선 인터넷 서비스는 구조 요청, 범죄 신고에의 대응, 인접 지역 정보 제공의 지리 정보 시스템(GIS: Geographic Information System), 위치에 따른 이동 통신 요금의 차등화, 교통 정보, 차량 항법 및 물류 관제, 위치 기반 CRM(Customer Relationship Management) 등 다양한 분야 및 상황에 사용될 수 있다. 이러한 위치 기반 서비스를 이용하기 위해서는 이동통신 단말기의 위치를 파악하는 것이 필수적이다. 이동통신 단말기의 위치를 파악하는 기술을 무선 측위 기술(PDT: Position Determination Technology)이라고 하는데, 기지국 수신 신호를 이용하는 망 기반(Network-Based) 방식과 GPS(Global Positioning System) 신호를 이용하는 핸드셋 기반(Handset-Based) 방식으로 구별되며, 최근에는 두 가지 기술을 혼합하여 위치 정확도를 높이는 하이브리드(Hybrid) 방식의 기술이 개발되고 있다. GPS 신호를 이용하는 핸드셋 기반 방식은 GPS 신호를 이용함으로써, 측위 정확도가 높다는 장점이 있으나, 이동통신 단말기에서 GPS 모듈을 이용하여 위치 정보를 수집하고, 수집한 위치 정보를 위치 결정 서버로 전송하여 이동통신 단말기의 위도 및 경도 좌표를 산출해야 하므로 이동통신망의 부하가 커지게 되며, 연속적인 위치 기반 서비스를 이용하기 어렵다는 단점이 있다. 또한, 특정 건물의 내부에 진입하는 경우 GPS 신호가 수신되지 않는다는 단점이 있으며, 기본적으로 GPS 모듈을 탑재한 이동통신 단말기의 가격이 비싸다는 단점이 있다. 기지국으로부터 수신되는 신호를 이용하는 망 기반 방식은 None GPS 이동통신 단말기에서 다수 개의 기지국으로부터 수신되는 신호를 이용하여 이동통신 단말기의 위치를 측위함으로써, 연속적인 위치 기반 서비스를 제공받을 수 있으며, 특정 건물의 내부에 진입한 경우에도 위치 기반 서비스를 제공받을 수 있다는 장점이 있다. 또한, 망 기반 방식은 GPS 모듈을 구비하지 않은 모든 기종의 단말기에 적용할 수 있는 장점이 있다. 하지만, 망 기반 방식은 이동통신 단말기와 기지국 간에 발생하는 전파 다중경로 현상으로 인해 이동통신 단말기의 위치를 측위하는 과정에서 측위 정확도가 떨어지는 단점이 있다. 특히, 망 기반 방식은 이동통신망에 광중계기가 포함된 경우 지연측정 오류를 발생하여 위치 측위의 정확도가 현격하게 떨어져서 실질적인 사용이 어렵다는 단점이 있다. 한편, 최근에는 특정 서버에서 GPS 모듈을



C:\Aviv\Patents\DEMO\DataTest\Korean_P_1020070103419.txt

광중계기 (I)
파라미터 (I)
기지국(122)과 (I)
광중계기를 (I)
광케이블과 (I)
소자(129)는 (I)
4km) (I)
기지국(122)으로부터 (I)
보유한다. (I)
None (I)
구현하며 (I)
기지국(122)이 (I)
산출하는 (I)
이하인 (I)
기지국(122)은 (I)
연동하도록 (I)
송신기는 (I)
Controller)가 (I)
광중계기(128)에 (I)
광중계기: (I)
광케이블의 (I)
usec (I)
있어서는 (I)

I: 326 E:553

Seq. ViewContents.asp (E)
이야기 (E)
Idx (E)
오신것을 (E)
Map (E)
Design (E)
Software, Mapping (E)
Music (E)
모바일 (E)
Asp • Cached page (E)

I: 17 E:553

의 (I)
기지국(122)의 (I)
주축하며 (I)
)은 (I)
이야기 (E)
Mobile (E)

I: 17 E:19

Chinese_P_200510002264.txt
Korean_P_1020070103419.txt
Korean_P_1020080110680.txt
Korean_P_1020080112065.txt
Korean_P_1020080115381.txt
Korean_P_1020080116416.txt
Korean_P_1020090037632.txt
Korean_P_1020097009627.txt
Korean_P_1020097009628.txt
Korean_P_1020097010714.txt
Korean_P_1020097011553.txt
Korean_P_1020097013795.txt
Korean_P_1020097014379.txt
Korean_P_1020097014576.txt
Korean_P_1020097015529.txt
Korean_P_1020097015530.txt

AST / α-I / α-E / no of docs: 36 / I 0.47 / I 0.25 / 38

Relevant Patents
Multiple Languages

Autom
keywords

User selects
elevant words

Multilingual words
selected

Korean

본 발명은 피셀 데이터베이스를 이용한 위성 항법 시스템을 구비하지
않은 이동통신 단말기의 위치 측정 방법, 서버 및 시스템에 관한 것이다.
본 발명은 파일롯 세기 측정 메시지(PSMM: Pilot Strength
Measurement Message, 이하 'PSMM'이라 칭함)를 생성하여 전달하는
이동통신 단말기; 이동통신 단말기로 이동통신 서비스를 제공하는
이동통신망; 이동통신 단말기로부터 수신하는 PSMM에 포함된 의사
잡음(PN: Pseudo Noise, 이하 'PN'이라 칭함)을 추출하여 PN의 개수 및
피셀 데이터베이스(Pilot Cell Database)의 존재 여부에 따라 특정 위치
측정 방식을 결정하고, 특정 위치 측정 방식에 따라 이동통신 단말기의
위치를 측정하여 위치 측정 결과를 전달하는 위치 계산 서버; 피셀
데이터베이스를 구비하고, 위치 계산 서버로부터 이동통신 단
위치 측정을 요청받으면 피셀 데이터베이스를 이용하여 이동
단말기의 위치를 측정하여 위치 계산 서버로 전달하는 피셀
서버; 및 위치 계산 서버로부터

Chinese

本发明涉及一种利用ＧＰＳ（Ｇｌｏｂａｌ　Ｐｏｓｉｔｉｏｎｉ
ｎｇ　Ｓｙｓｔｅｍ：ＧＰＳ）的移动通讯终端的位置追踪方法，
其特征在于，包括：用户根据菜单来设置ＧＰＳ－ＯＮＥ操作维持功
能活性化；**判断是否将ＧＰＳ－ＯＮＥ操作**维持功能活性化（ｅｎａ
ｂｌｅ）的阶段；其判断结果为已确认上述ＧＰＳ－ＯＮＥ操作维持
功能的活性化时，即使合上移动通讯终端翻盖或按下结束键也仍然能
维持ＧＰＳ－ＯＮＥ操作的控制阶段；否则，中断ＧＰＳ操作
明可防止用户因过失而折叠翻盖按下结束键引发的ＧＰＳ操作
主权利要求：
一种利用ＧＰＳ的移动通讯终端的位置追踪方法，其特
用户根据菜单来设置ＧＰＳ－ＯＮＥ操作维持功能活
将ＧＰＳ－ＯＮＥ操作维持功能活性化的阶段；其判
上述ＧＰＳ－ＯＮＥ操作维持功能的活性化时，即使
端翻盖或按下结束键也仍然能维持ＧＰＳ－ＯＮＥ操

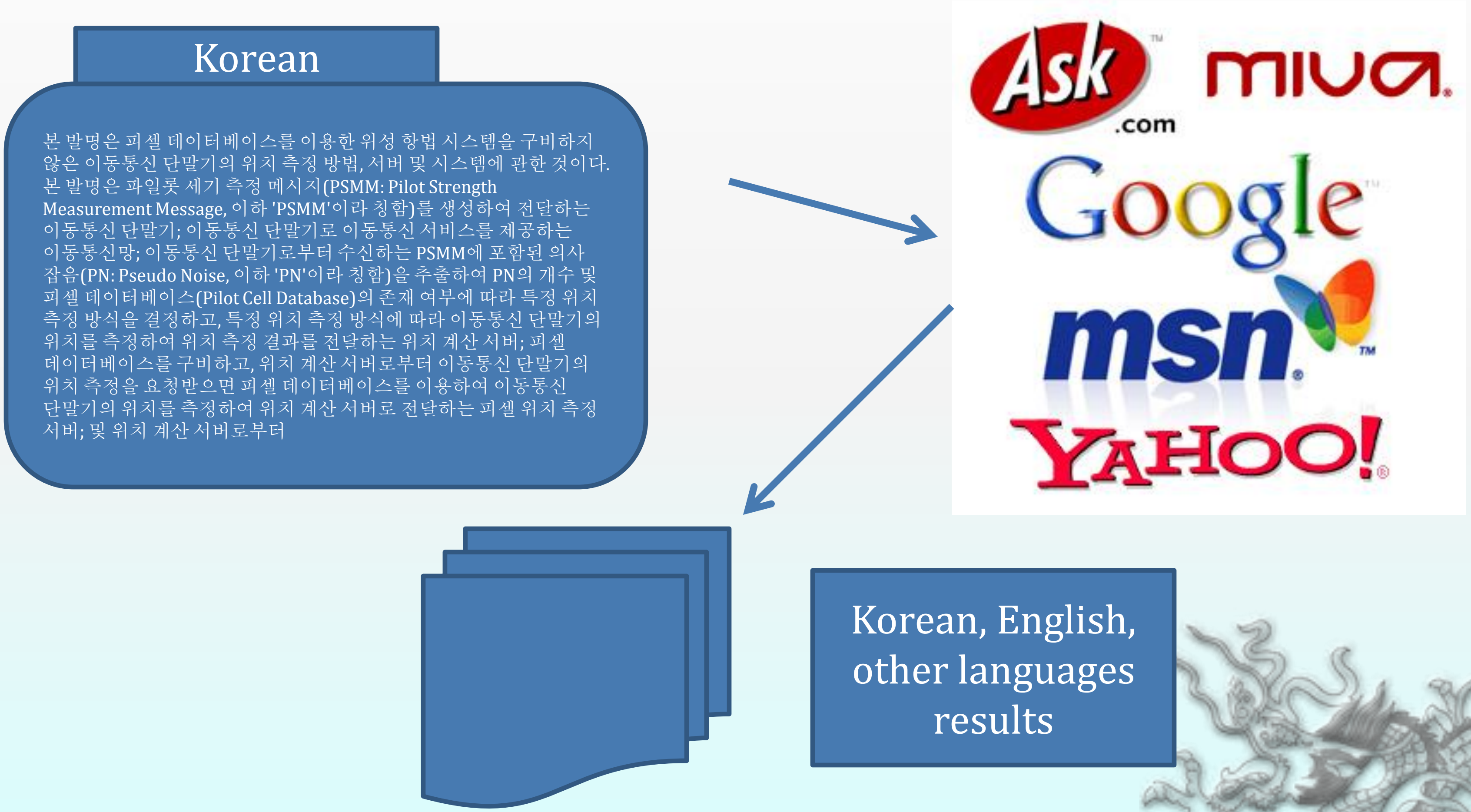
English

Signal comparison-based location determining method

Abstract
At least one portable RF communications device in conjunction with at least two fixed-location service-area antenna stations respectively capable of RF communication with the at least one device performs the steps of: (I) using a portable device at a selected location to measure RF communications signals from the plurality of local fixed-location service-area antenna stations and electronically storing at least two of the respective reception signal strength measurements; and (II) monitoring a portable device location by causing the device to measure reception signal strength associated with local fixed-location service-area antenna stations signals, and to electronically compare these measurements with the stored at least two measurements.

Multilingual Model

- Web search



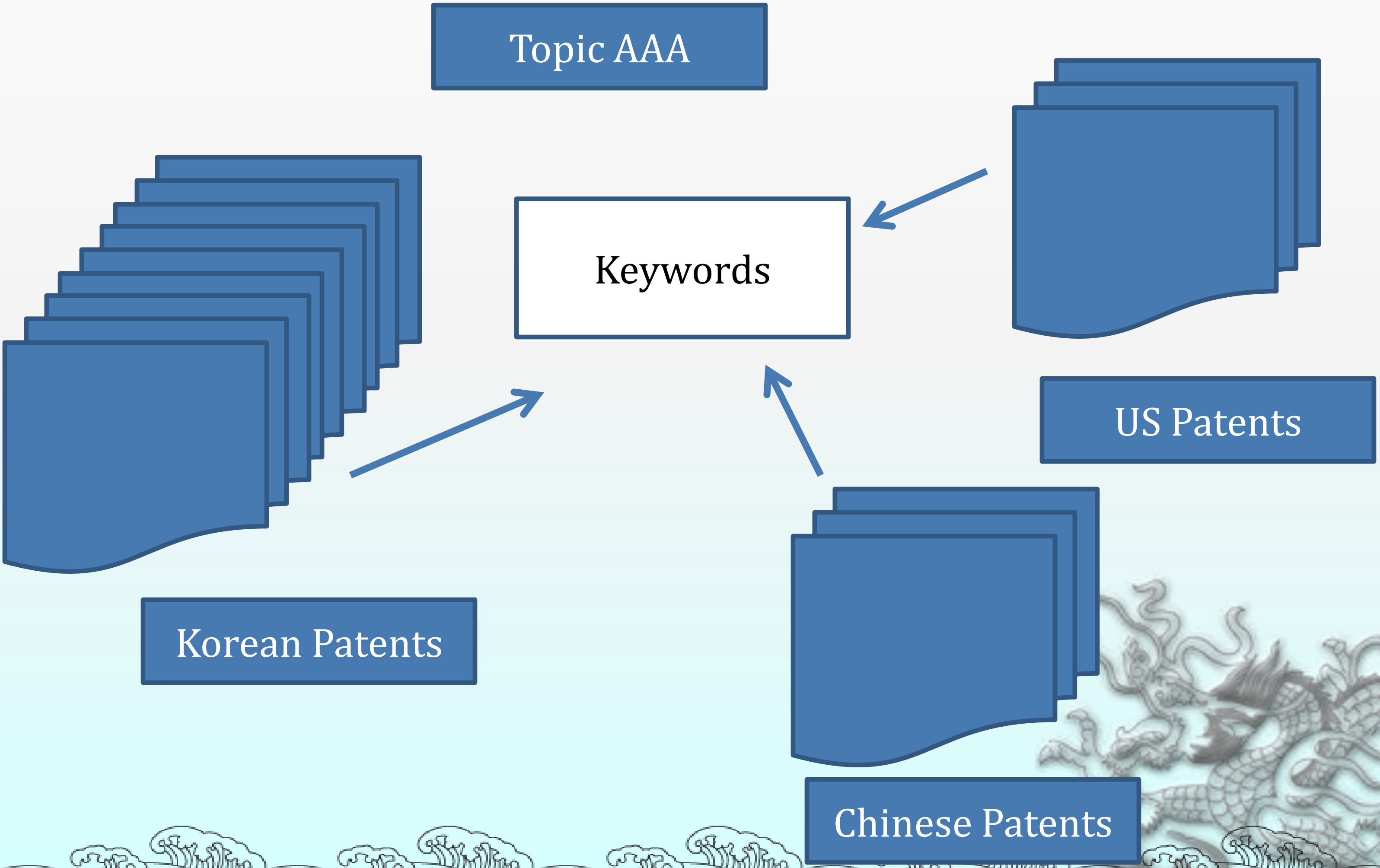
Korean

본 발명은 피셀 데이터베이스를 이용한 위성 항법 시스템을 구비하지 않은 이동통신 단말기의 위치 측정 방법, 서버 및 시스템에 관한 것이다. 본 발명은 파일롯 세기 측정 메시지(PSMM: Pilot Strength Measurement Message, 이하 'PSMM'이라 칭함)를 생성하여 전달하는 이동통신 단말기; 이동통신 단말기로 이동통신 서비스를 제공하는 이동통신망; 이동통신 단말기로부터 수신하는 PSMM에 포함된 의사 잡음(PN: Pseudo Noise, 이하 'PN'이라 칭함)을 추출하여 PN의 개수 및 피셀 데이터베이스(Pilot Cell Database)의 존재 여부에 따라 특정 위치 측정 방식을 결정하고, 특정 위치 측정 방식에 따라 이동통신 단말기의 위치를 측정하여 위치 측정 결과를 전달하는 위치 계산 서버; 피셀 데이터베이스를 구비하고, 위치 계산 서버로부터 이동통신 단말기의 위치 측정을 요청받으면 피셀 데이터베이스를 이용하여 이동통신 단말기의 위치를 측정하여 위치 계산 서버로 전달하는 피셀 위치 측정 서버; 및 위치 계산 서버로부터

Korean, English, other languages results

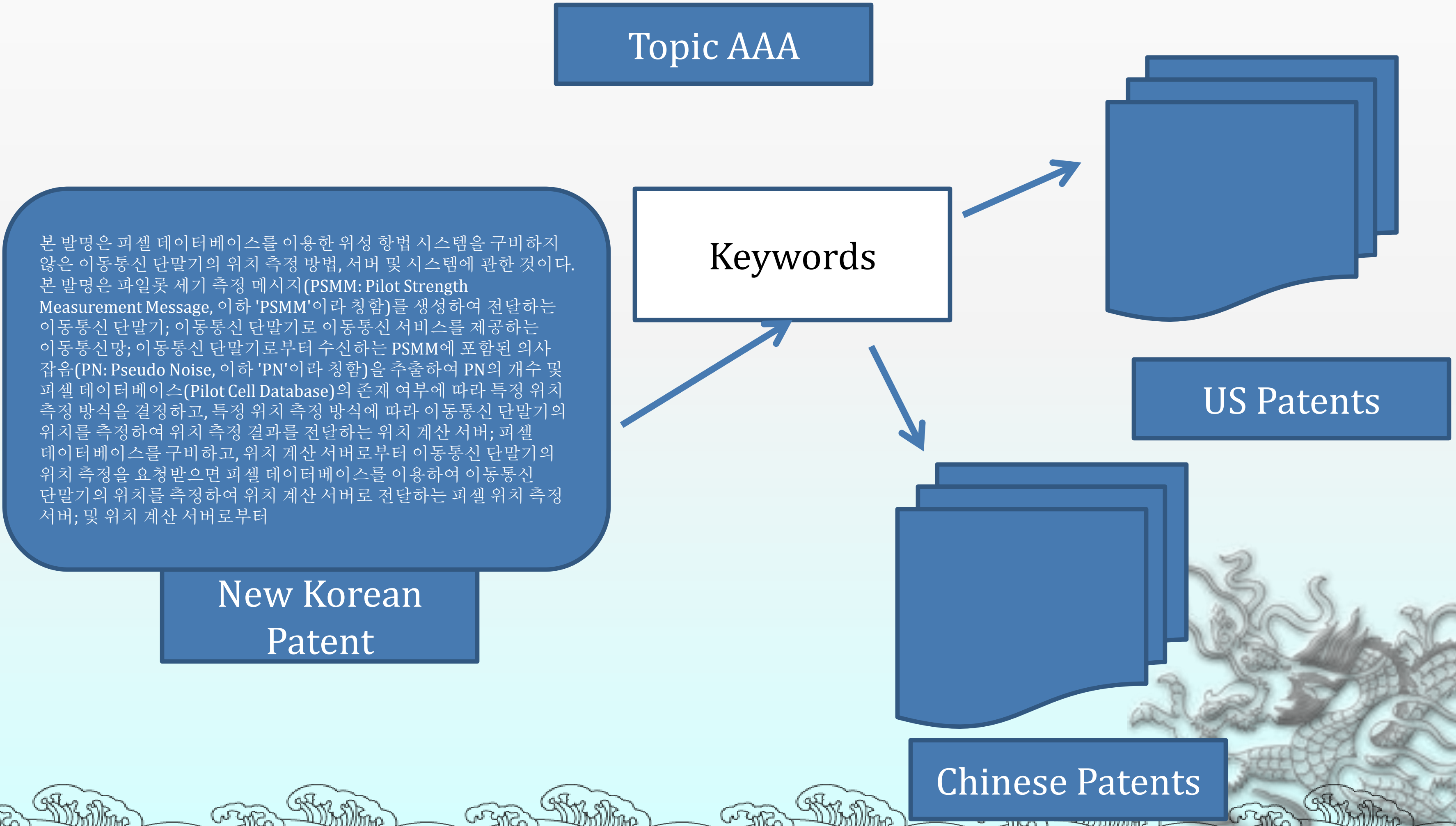
Multilingual Model

- Patent offices topics related patents



Multilingual Model

- Patent offices topics related patents



Topic AAA

본 발명은 피셀 데이터베이스를 이용한 위성 항법 시스템을 구비하지 않은 이동통신 단말기의 위치 측정 방법, 서버 및 시스템에 관한 것이다. 본 발명은 파일롯 세기 측정 메시지(PSMM: Pilot Strength Measurement Message, 이하 'PSMM'이라 칭함)를 생성하여 전달하는 이동통신 단말기; 이동통신 단말기로 이동통신 서비스를 제공하는 이동통신망; 이동통신 단말기로부터 수신하는 PSMM에 포함된 의사 잡음(PN: Pseudo Noise, 이하 'PN'이라 칭함)을 추출하여 PN의 개수 및 피셀 데이터베이스(Pilot Cell Database)의 존재 여부에 따라 특정 위치 측정 방식을 결정하고, 특정 위치 측정 방식에 따라 이동통신 단말기의 위치를 측정하여 위치 측정 결과를 전달하는 위치 계산 서버; 피셀 데이터베이스를 구비하고, 위치 계산 서버로부터 이동통신 단말기의 위치 측정을 요청받으면 피셀 데이터베이스를 이용하여 이동통신 단말기의 위치를 측정하여 위치 계산 서버로 전달하는 피셀 위치 측정 서버; 및 위치 계산 서버로부터

New Korean Patent

Keywords

US Patents

Chinese Patents

Prediction

It's tough to make predictions, especially about the future

(Markus M. Ronner, 1918)

Patent Trends – Predicting New Technologies

- 4,354,054 patents from the US Patent Office from 1975 until today
- Goals:
  - Find an equation that can predict technology/trend
  - Visualize change in technology/trend

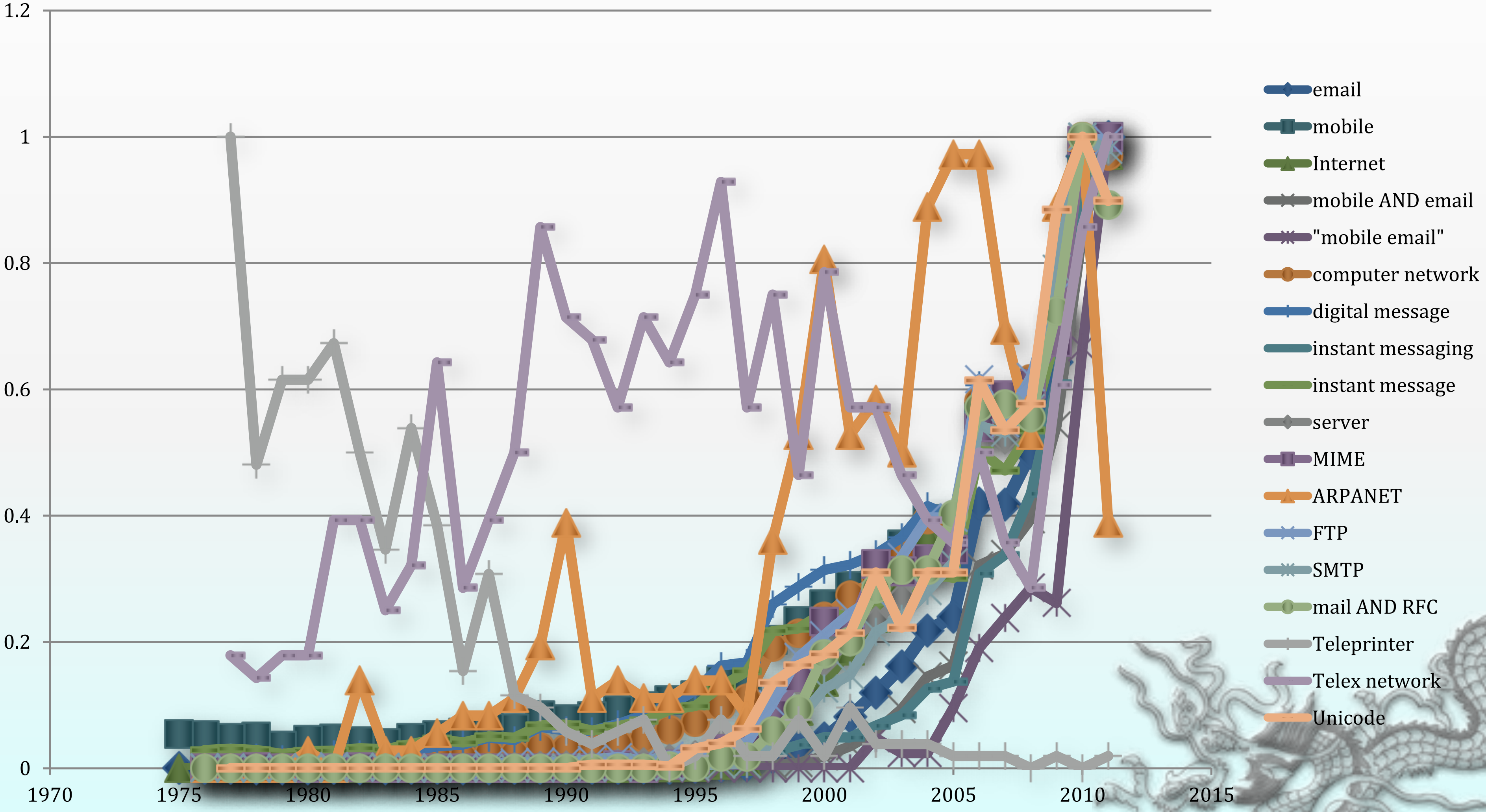
Technology Temporal Analysis Method

- ***Extracting Related Terms***
- ***Extracting All Graphs (term frequency)***
- ***Elimination Process***

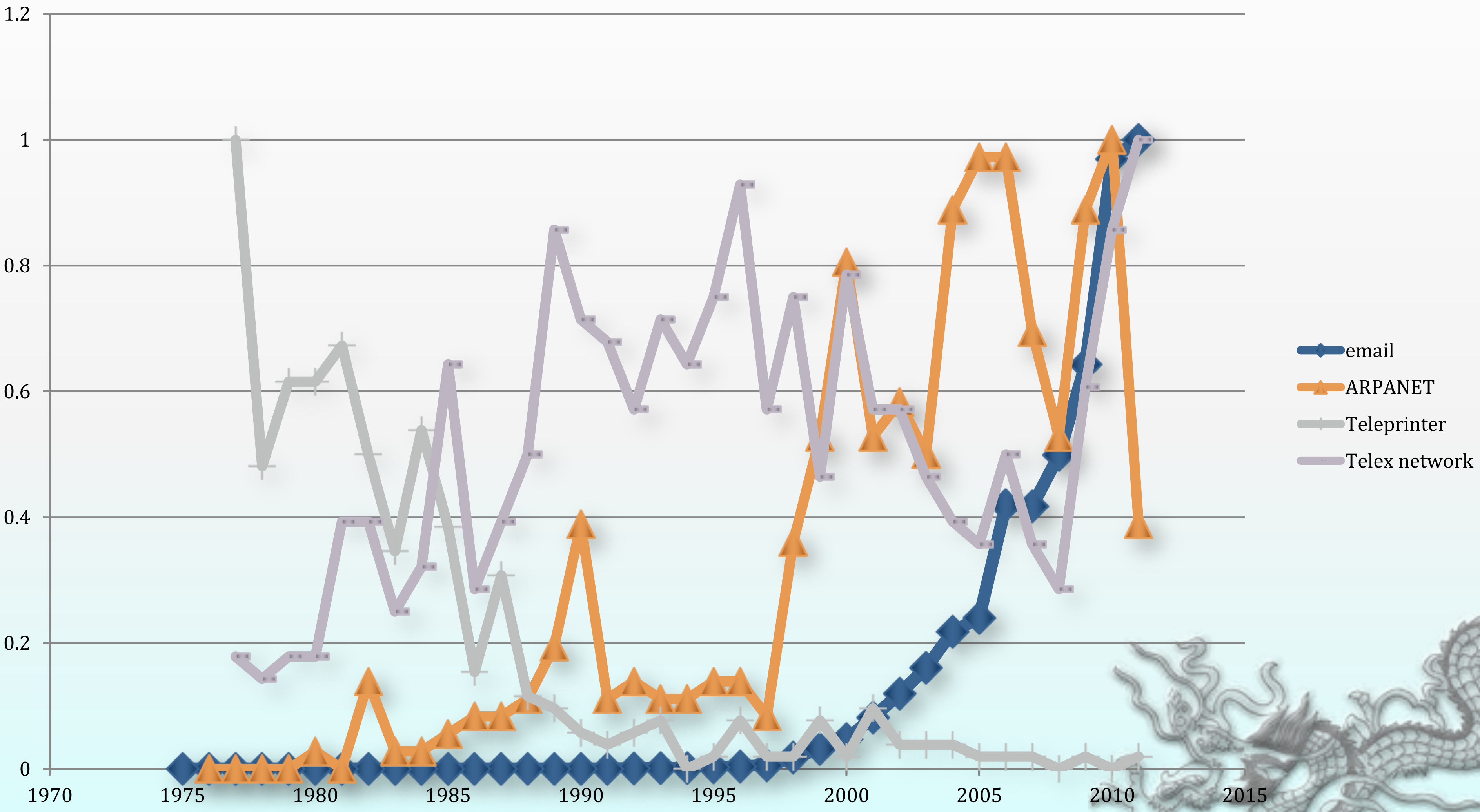
$$y= 0.055558046 * 1.160450815^{x} - 0.084088217,\ R^2<0.94$$

- ***Graph Distance (Δt time difference)***

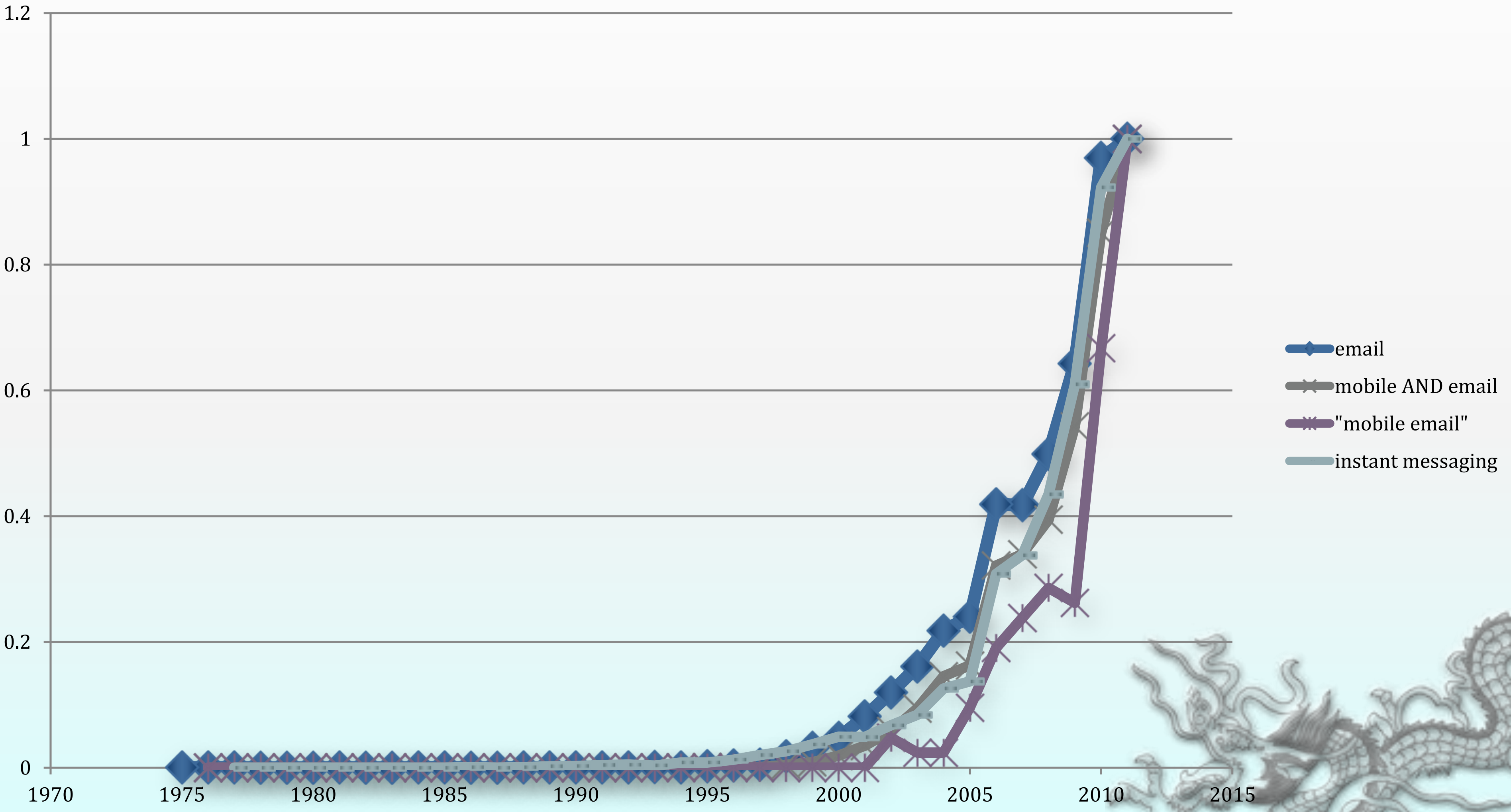
Patent Trend Prediction - email



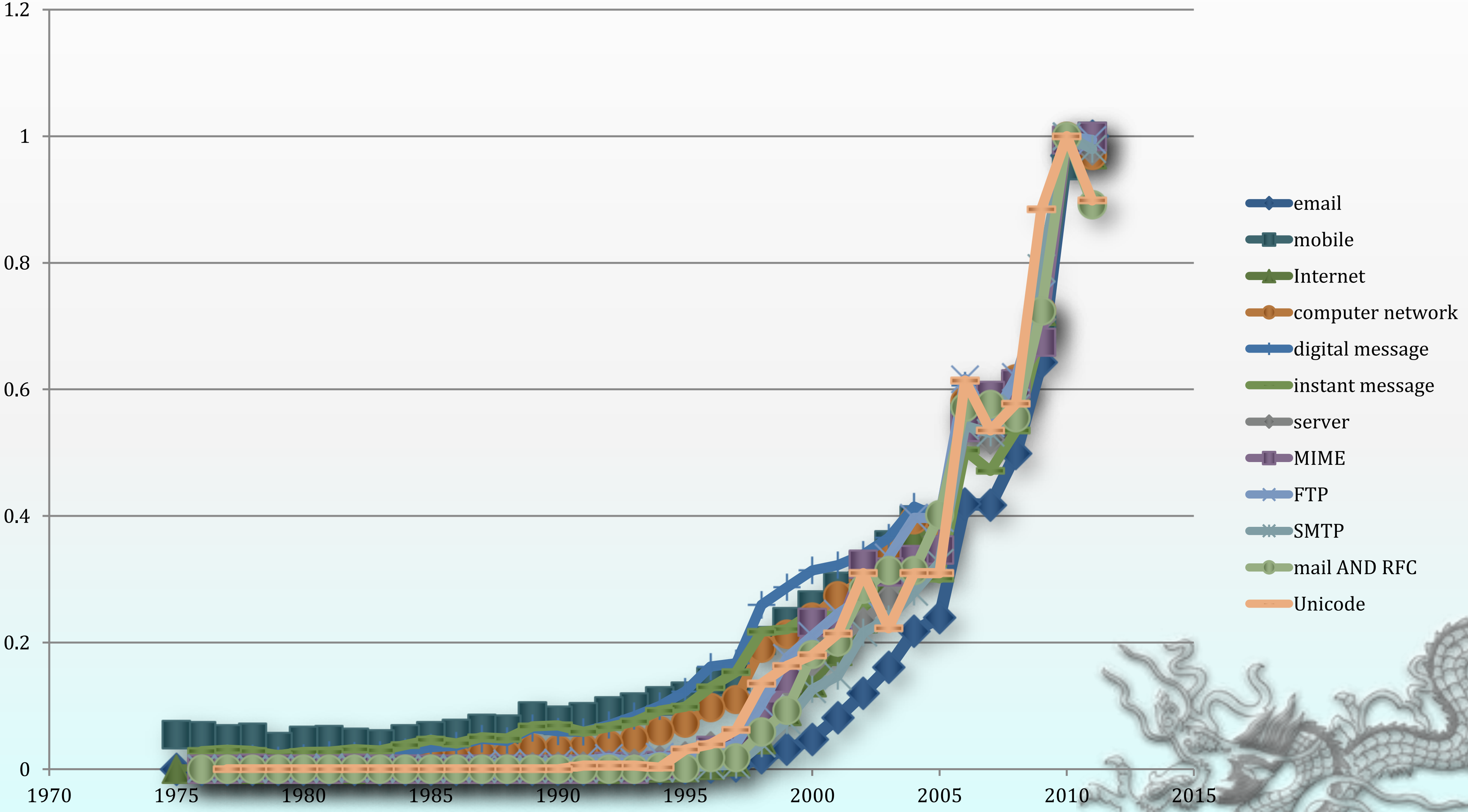
Patent Trend Prediction - email



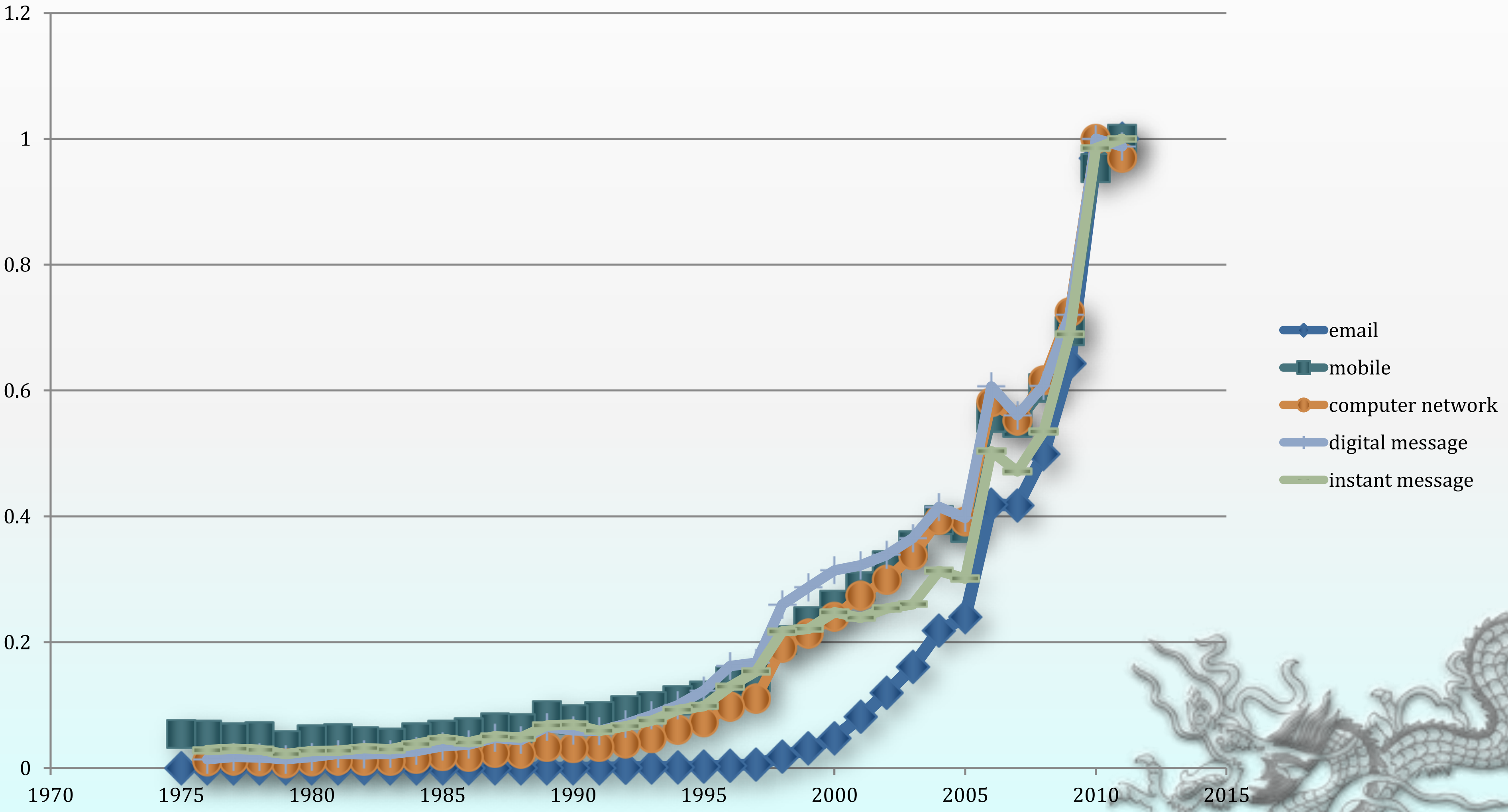
Patent Trend Prediction - email

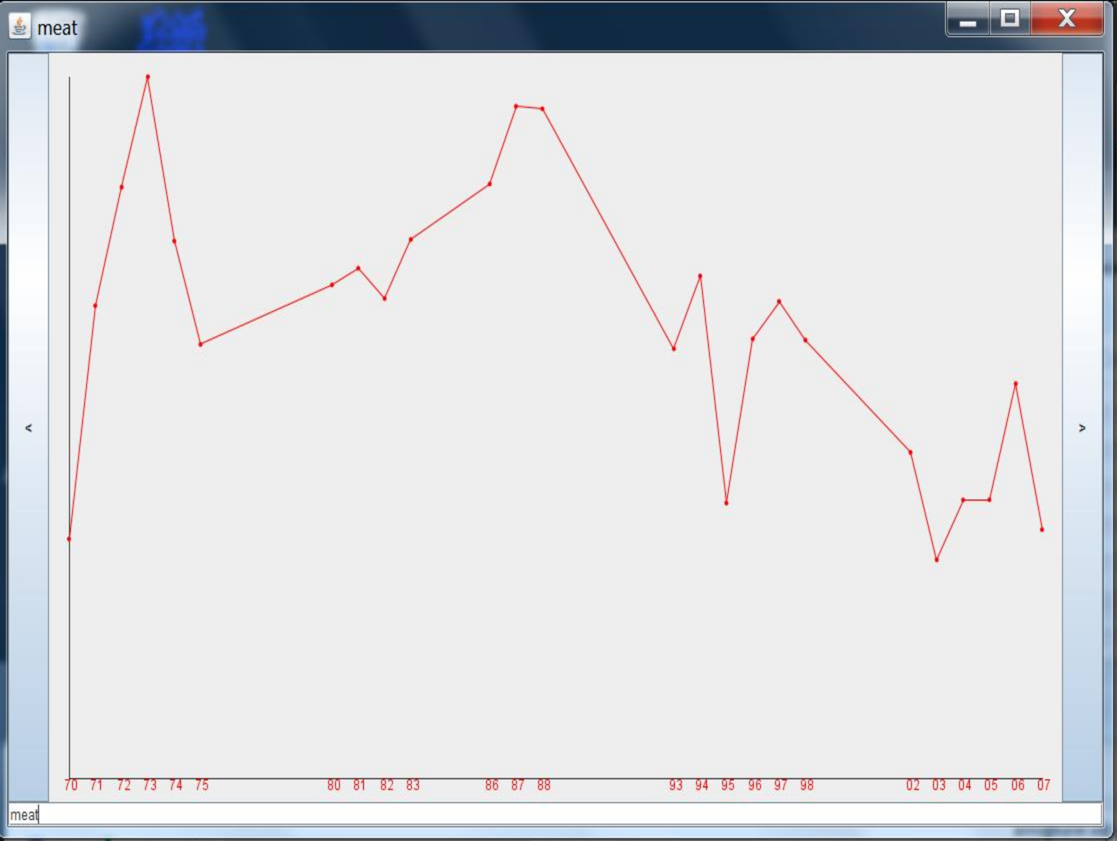


Patent Trend Prediction - email

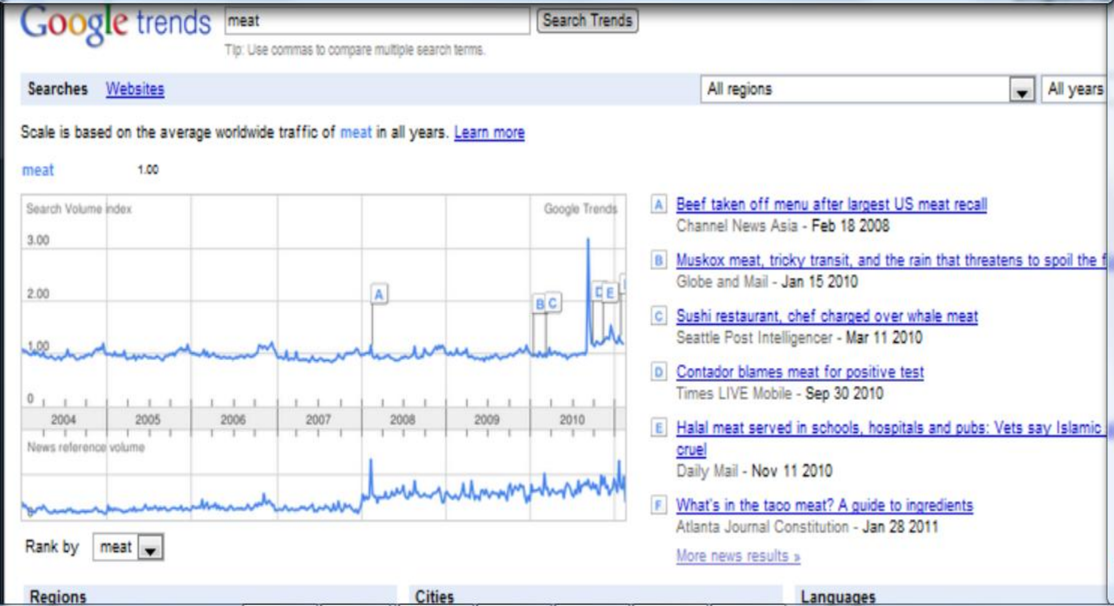


Patent Trend Prediction - email





meat



meat

Tip: Use commas to compare multiple search terms.

Searches Websites All regions All years

Scale is based on the average worldwide traffic of meat in all years. Learn more

meat 1.00

A Beef taken off menu after largest US meat recall
Channel News Asia - Feb 18 2008

B Muskox meat, tricky transit, and the rain that threatens to spoil the f
Globe and Mail - Jan 15 2010

C Sushi restaurant, chef charged over whale meat
Seattle Post Intelligencer - Mar 11 2010

D Contador blames meat for positive test
Times LIVE Mobile - Sep 30 2010

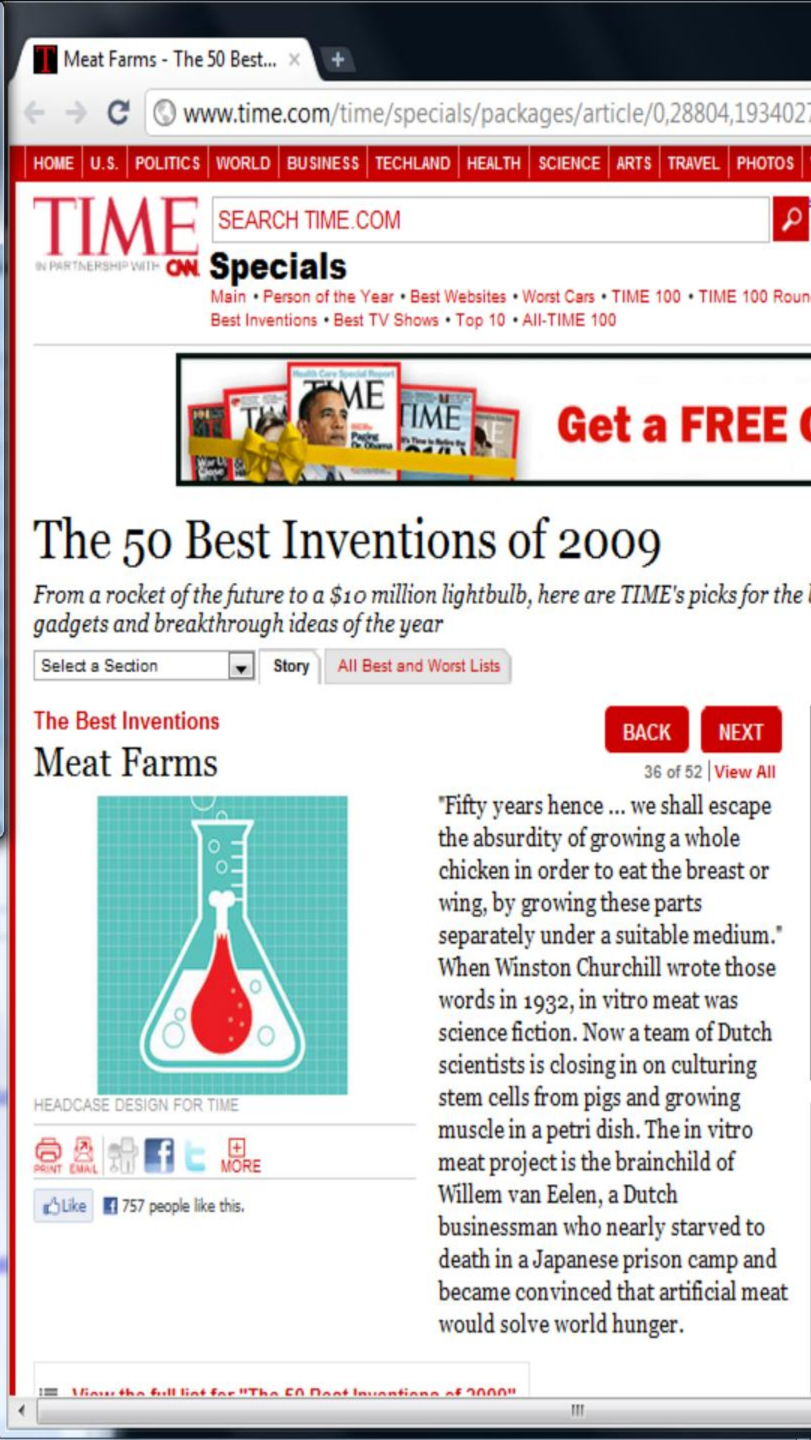
E Halal meat served in schools, hospitals and pubs: Vets say Islamic cruel
Daily Mail - Nov 11 2010

F What's in the taco meat? A guide to ingredients
Atlanta Journal Constitution - Jan 28 2011

More news results »

Rank by meat

Regions Cities Languages


Meat Farms - The 50 Best... ×
www.time.com/time/specials/packages/article/0,28804,193402
HOME U.S. POLITICS WORLD BUSINESS TECHLAND HEALTH SCIENCE ARTS TRAVEL PHOTOS
TIME
IN PARTNERSHIP WITH CNN
SEARCH TIME.COM
Get a FREE
PRINT EMAIL
MORE
Like
BACK
NEXT

Specials

Main • Person of the Year • Best Websites • Worst Cars • TIME 100 • TIME 100 Roun
Best Inventions • Best TV Shows • Top 10 • All-TIME 100

The 50 Best Inventions of 2009

From a rocket of the future to a $10 million lightbulb, here are TIME's picks for the gadgets and breakthrough ideas of the year

Select a Section | Story | All Best and Worst Lists

The Best Inventions

Meat Farms

36 of 52 | View All

HEADCASE DESIGN FOR TIME

757 people like this.

"Fifty years hence ... we shall escape the absurdity of growing a whole chicken in order to eat the breast or wing, by growing these parts separately under a suitable medium." When Winston Churchill wrote those words in 1932, in vitro meat was science fiction. Now a team of Dutch scientists is closing in on culturing stem cells from pigs and growing muscle in a petri dish. The in vitro meat project is the brainchild of Willem van Eelen, a Dutch businessman who nearly starved to death in a Japanese prison camp and became convinced that artificial meat would solve world hunger.

View the full list for "The 50 Best Inventions of 2009"

Patent Service
Self-Organizing Maps

Aviv Segev and Jussi Kantola, Identification of Trends from Patents Using Self-Organizing Maps, Journal of Expert Systems with Applications (ESWA), 39, pp. 13235–13242, 2012

Problem

- Identify the directions in which the new technology is advancing

- Government - Forecast main research areas that would be beneficial to fund

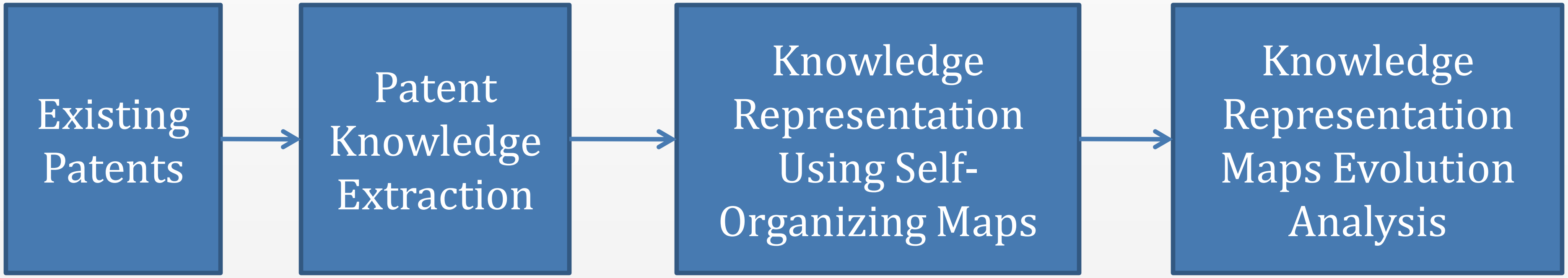
- Researchers - Map knowledge and identify possible gaps relevant to the advancement of science

Approach

- A model based on knowledge extraction from patents and self-organizing maps for knowledge representation

- The model was tested on patents from the United States Patent and Trademark Office.

PATENT KNOWLEDGE EXTRACTION

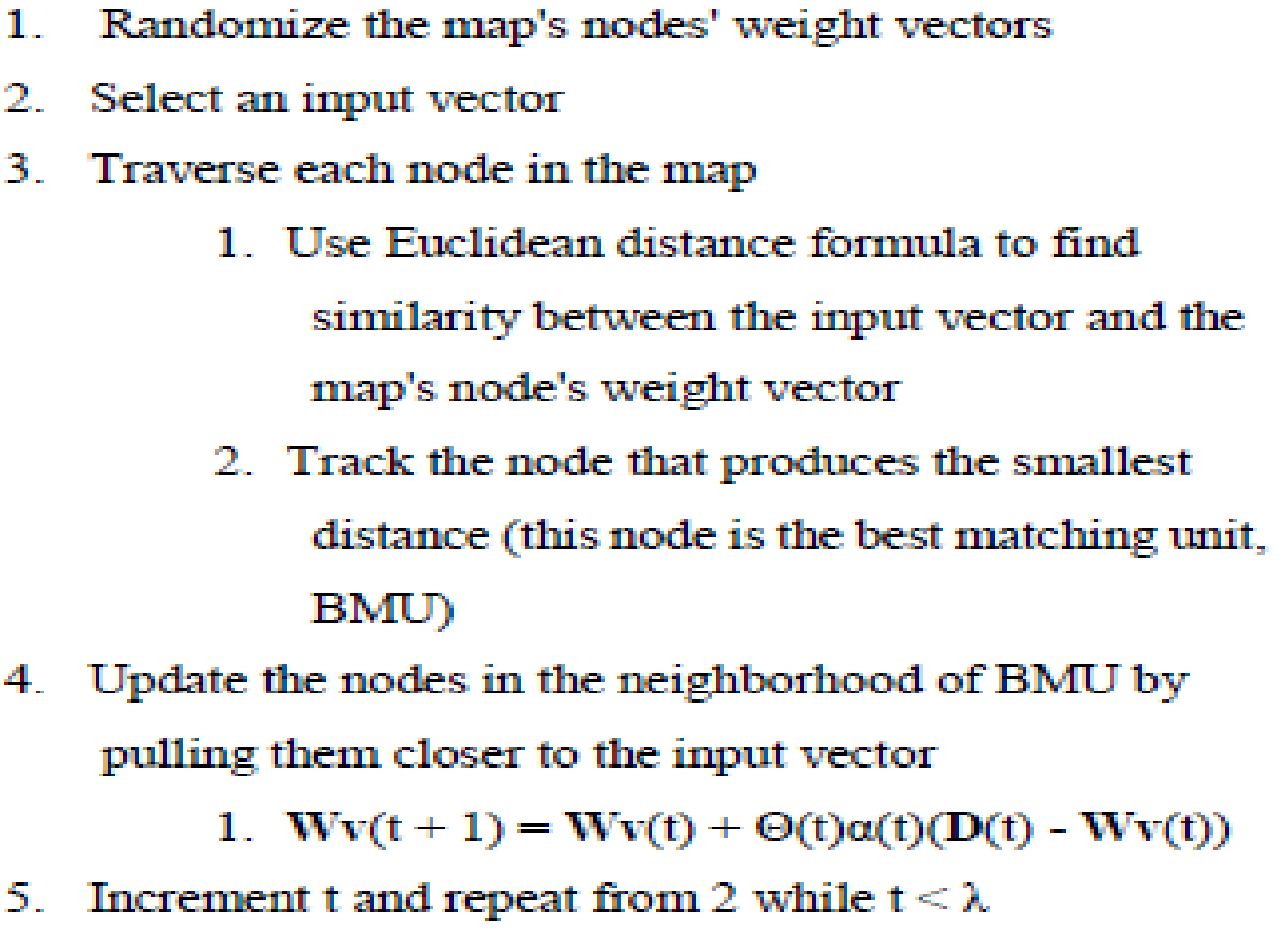

Existing Patents
Patent Knowledge Extraction
Knowledge Representation Using Self-Organizing Maps
Knowledge Representation Maps Evolution Analysis

- Context Extraction Using the Web
- Term Frequency / Inverse Document Frequency

SELF-ORGANIZING MAPS

- Self-Organizing Map (SOM) is a type of artificial neural network trained using unsupervised learning to produce a low-dimensional discretized representation of the input space of the training samples, called a map. (Kohonen, 2001)
- Self-Organizing Maps by key concepts in patents

SOM Learning Algorithm



1. Randomize the map's nodes' weight vectors
2. Select an input vector
3. Traverse each node in the map
    1. Use Euclidean distance formula to find similarity between the input vector and the map's node's weight vector
    2. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
4. Update the nodes in the neighborhood of BMU by pulling them closer to the input vector
    1. $\mathbf{Wv}(t + 1) = \mathbf{Wv}(t) + \Theta(t)\alpha(t)(\mathbf{D}(t) - \mathbf{Wv}(t))$
5. Increment t and repeat from 2 while $t < \lambda$

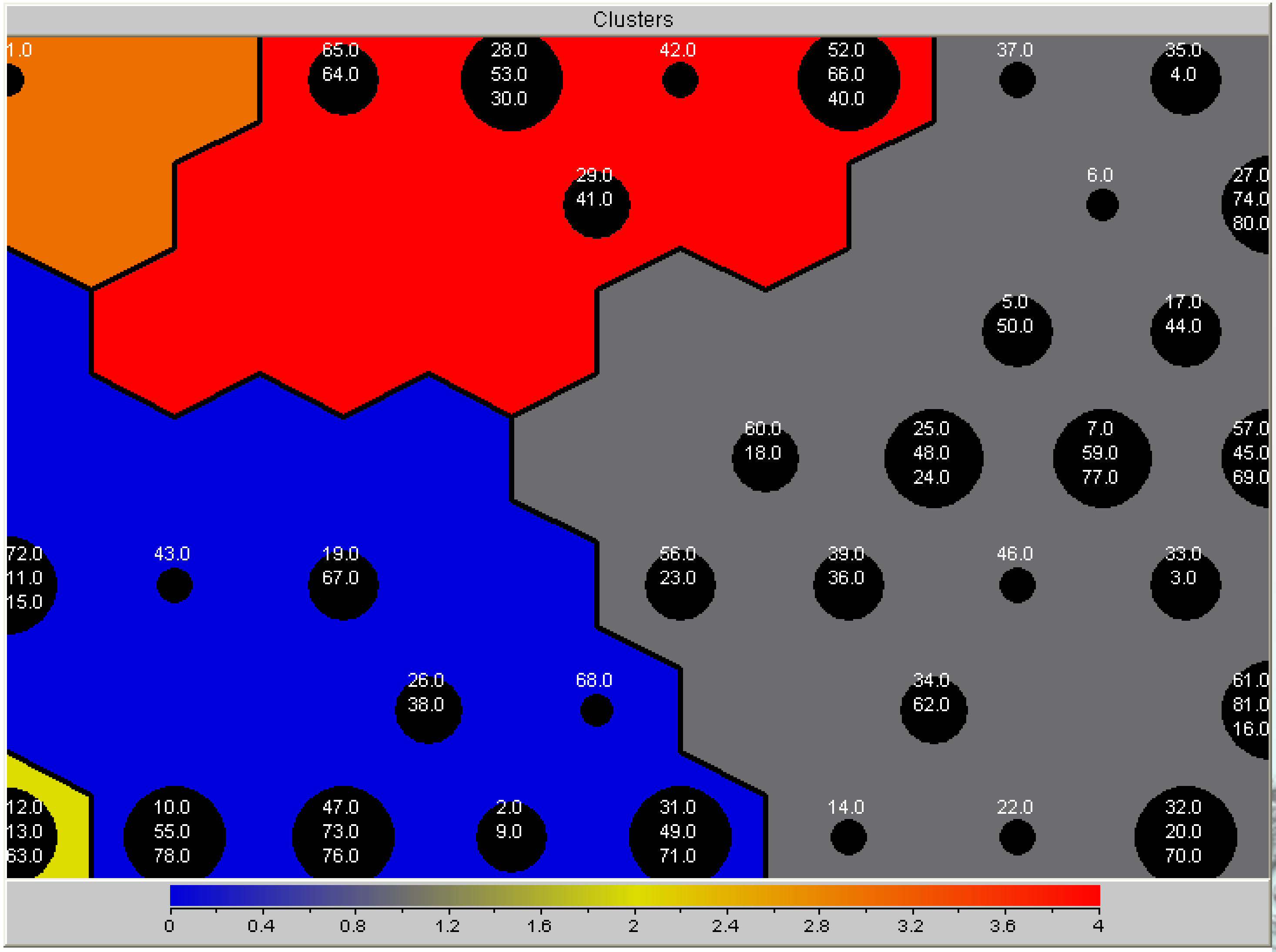
Experiments

- 81 patents from the United States Patent and Trademark Office
- 43 top ranking context values.

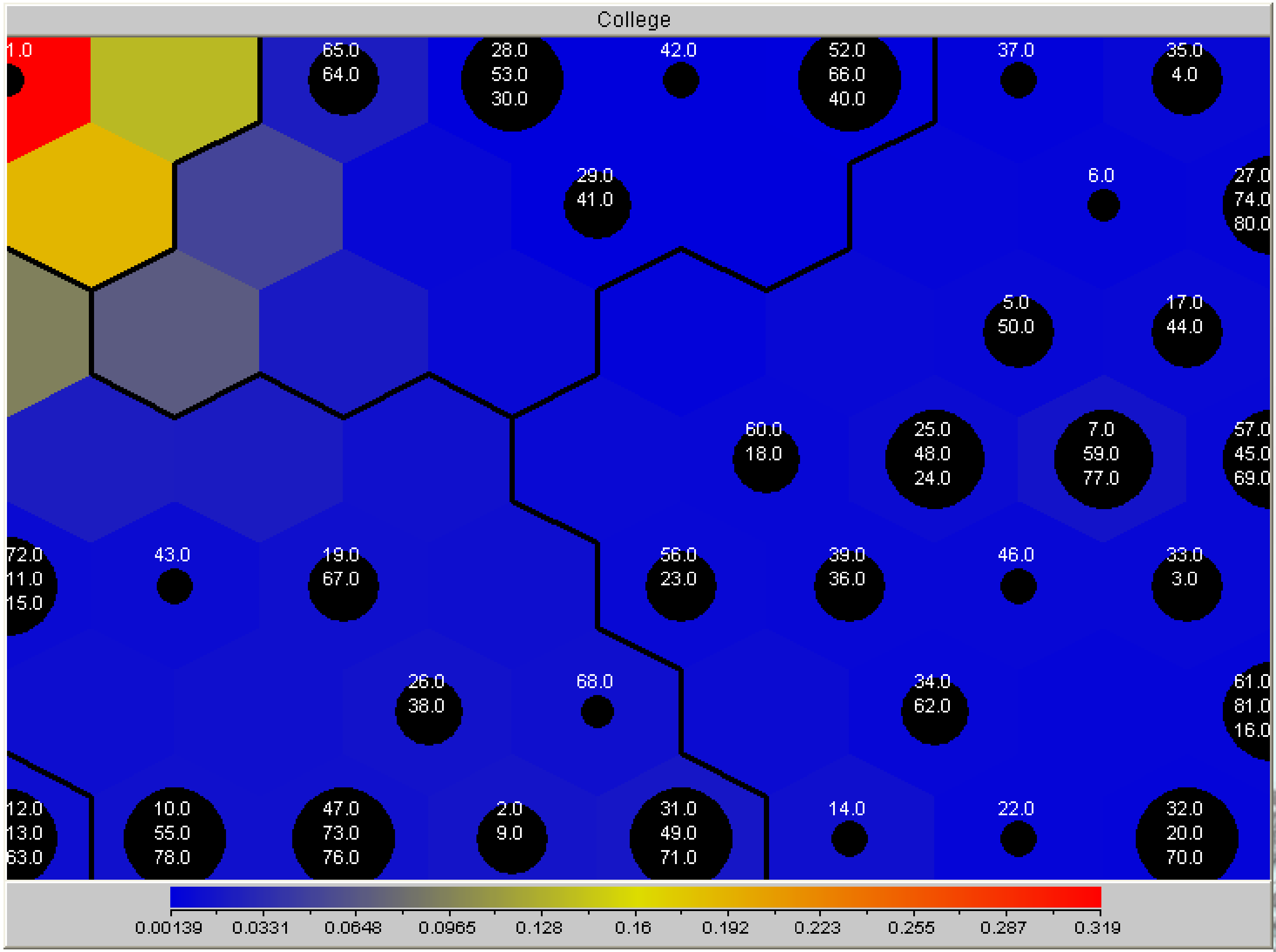
The experiments included:

- Identifying the main clusters of the patents.
- Analyzing the patent maps according to each context to identify meaningful contexts.

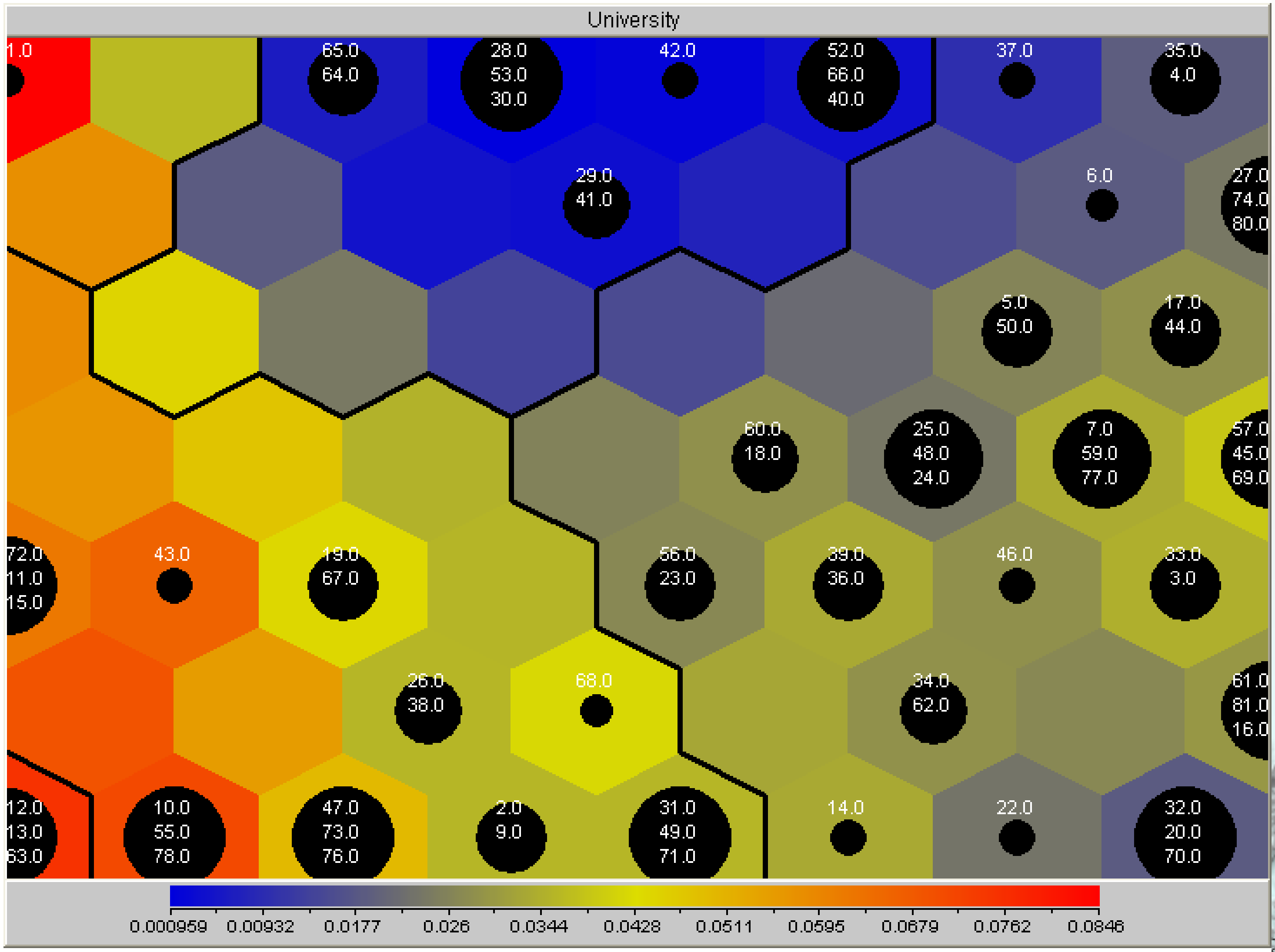
SOM Patent Clusters



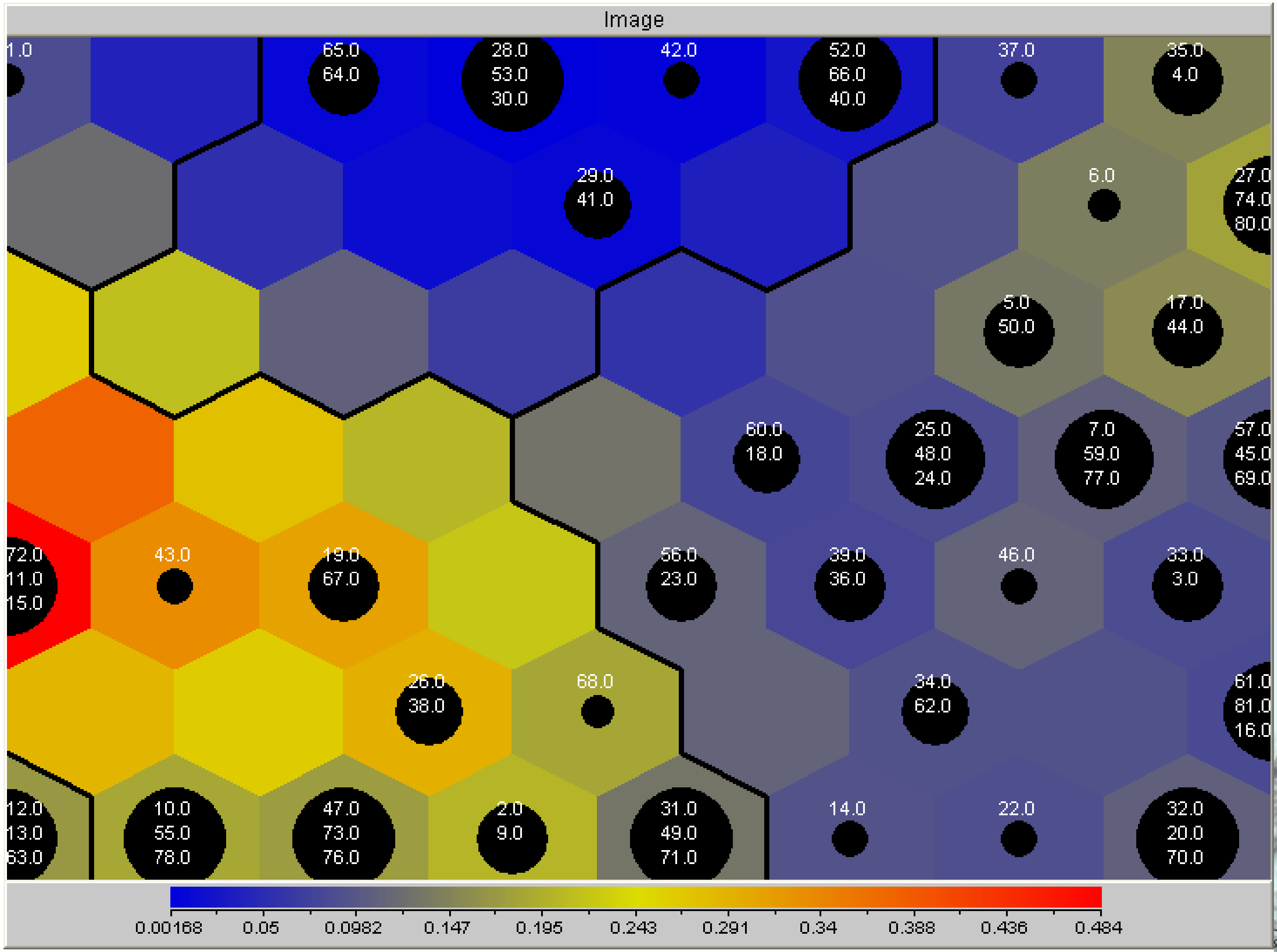
SOM College Patents



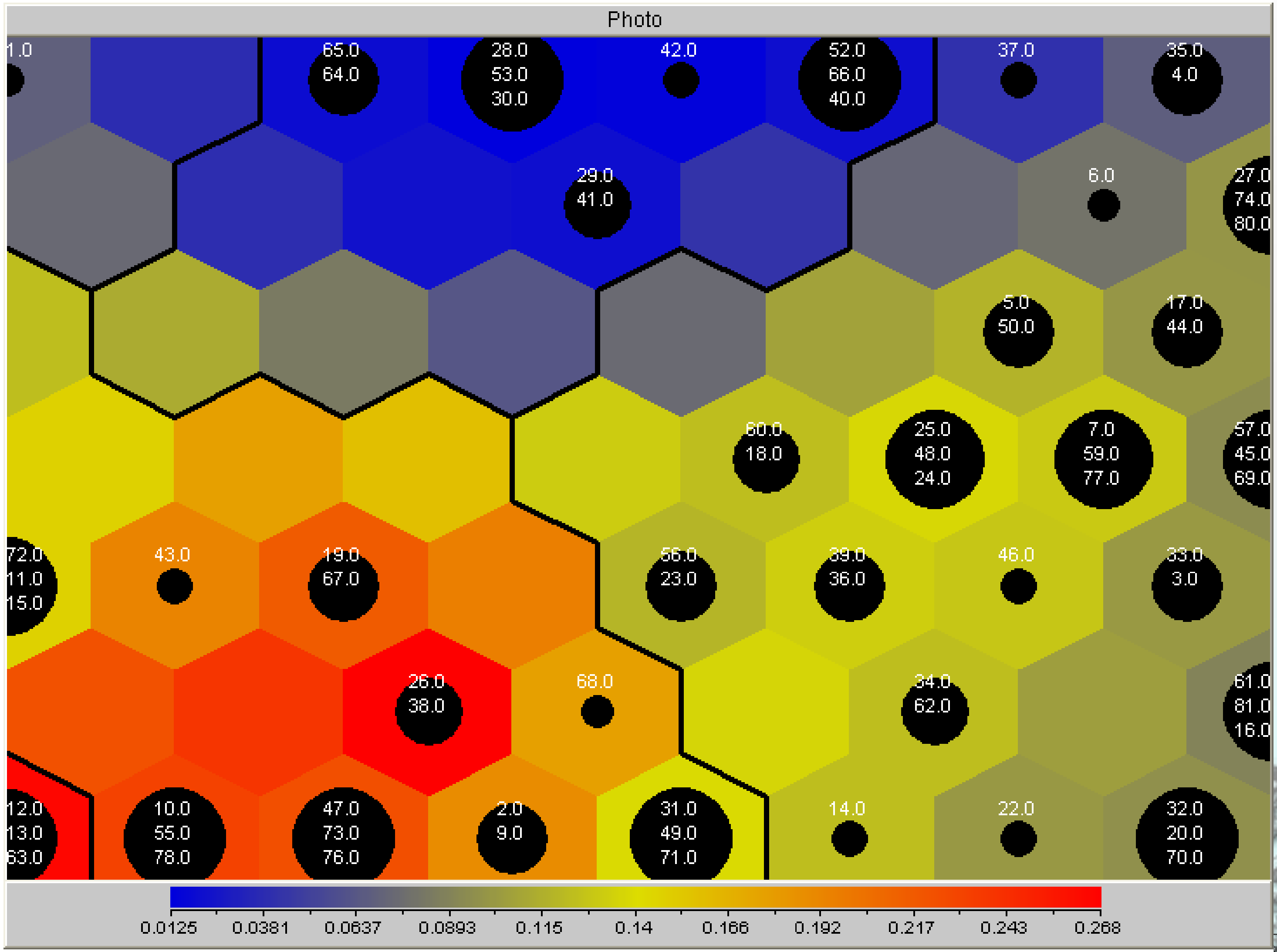
SOM University Patents



SOM Image Patents



SOM Photo Patents



Community Prediction in Citation Networks

Sukhwan Jung

- Purpose
  - To see if the structural information of a social network can be used to predict changes in the communities.
  - To test the citation networks as a dataset

Research Outline

- Gather data from Social Network and create a temporal map of concepts(communities) in certain domain, showing how concepts change over time.

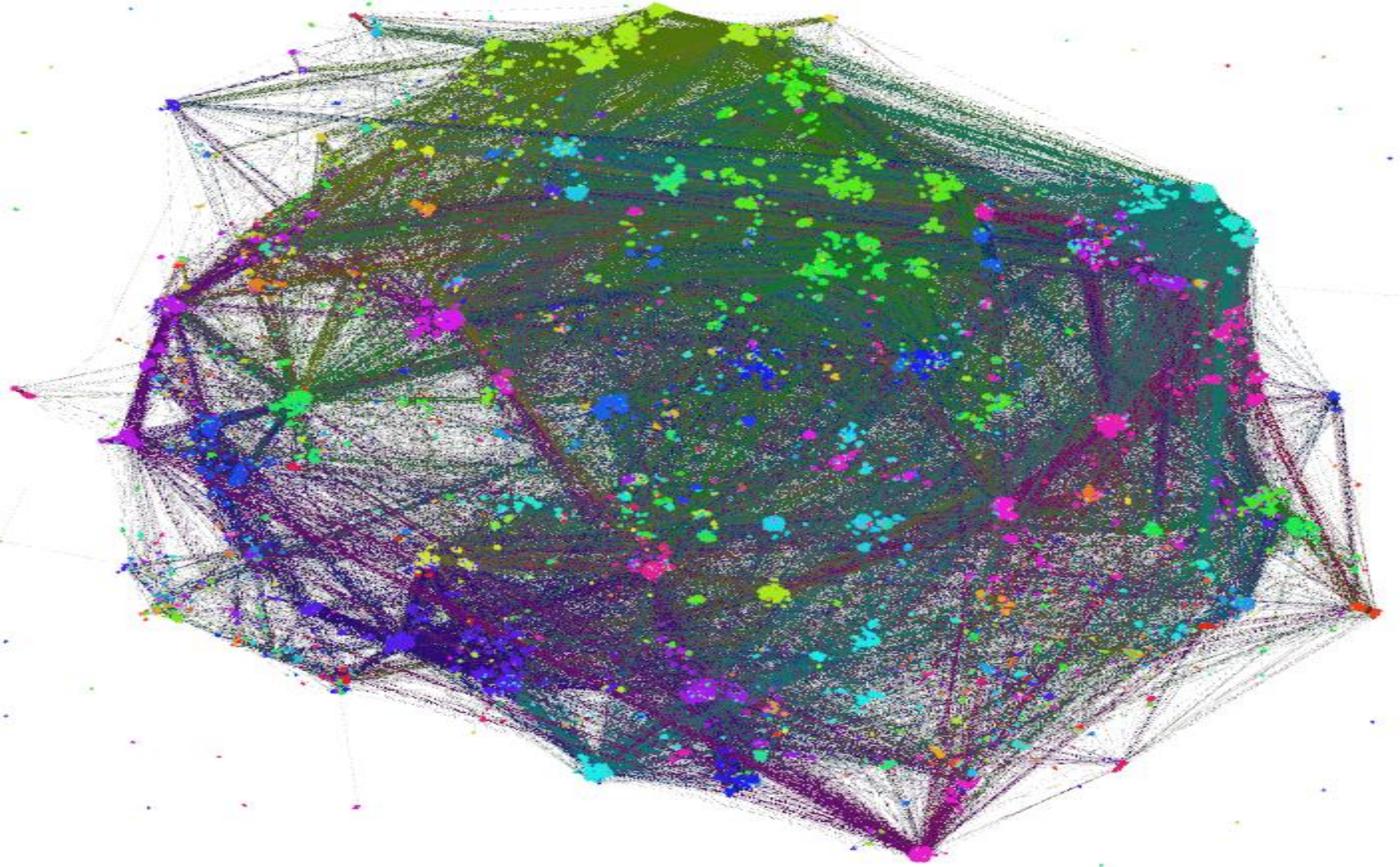
- Calculate the user movement over concept in the next timeline; predict the (*un*)*popular* concepts.

Data

- Citation network
  - Node = Research paper
  - Edge = Citation
  - Nodes & Edges do not disappear in citation network

  - High Energy Physics(hepPh)
    - 30566 papers, 347414 citations
  - High Energy Physics Theory(hepTh)
    - 18479 papers, 136428 citations

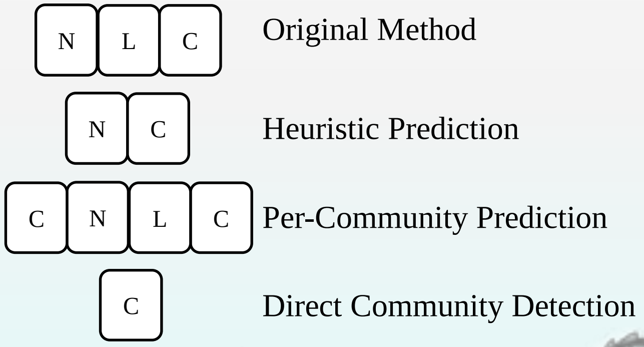
Result



Proposed Methods

3 modules used

- **N**ode prediction module
  - Newly proposed to predict how many nodes will appear in the future
- **L**ink prediction module
  - Existing link prediction methods
- **C**ommunity detection module
  - Existing community detection methods



N L C Original Method

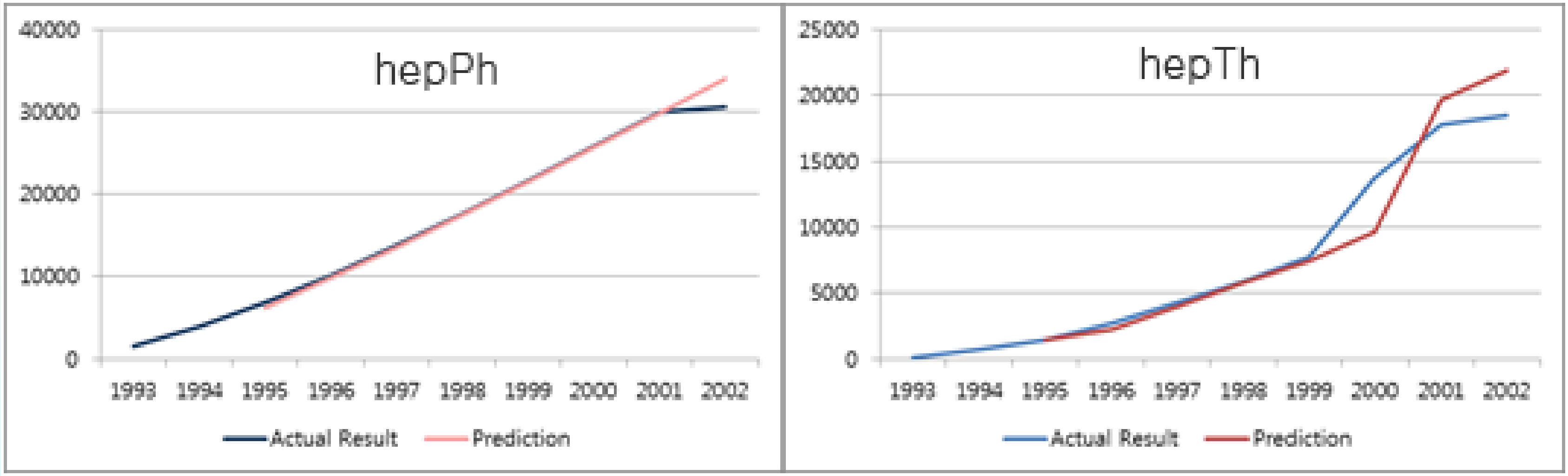
N C Heuristic Prediction

C N L C Per-Community Prediction

C Direct Community Detection

Node prediction result

- # nodes: correlation coefficient *r* = 0.98, 7.5% margin of error.

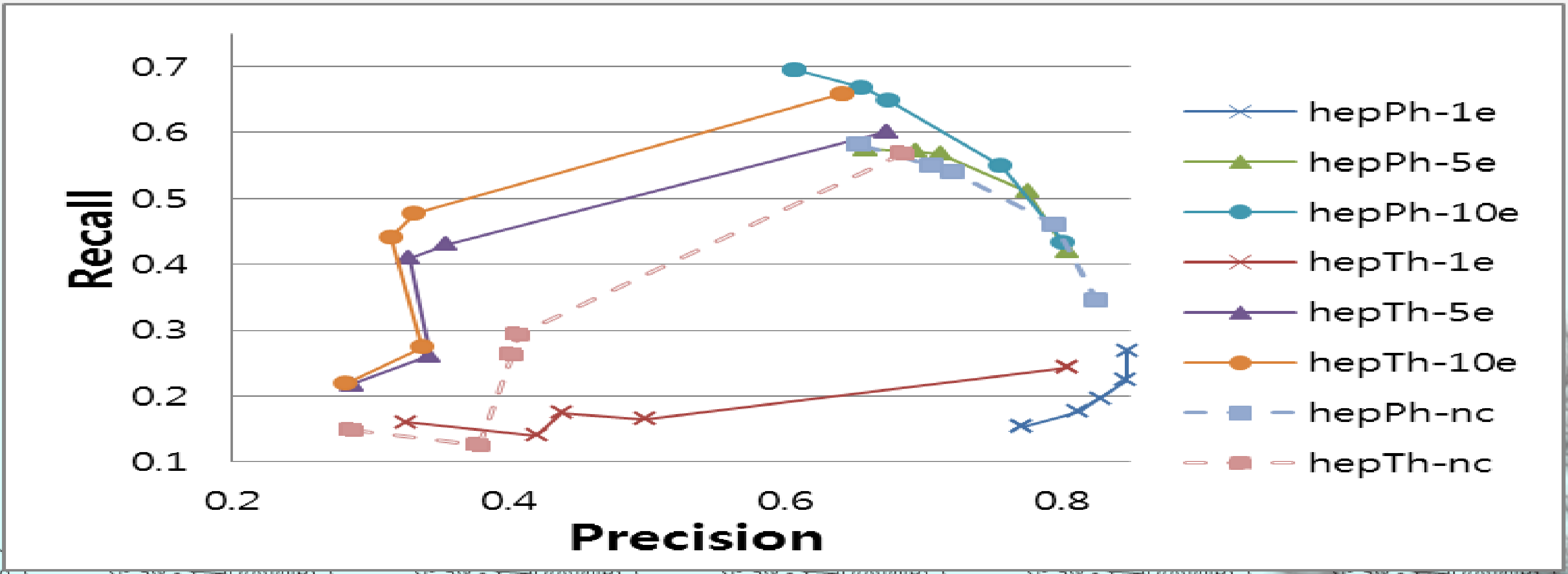


Results

- Edges are predicted by Node prediction module
  - *Number of edges to predict per node:*

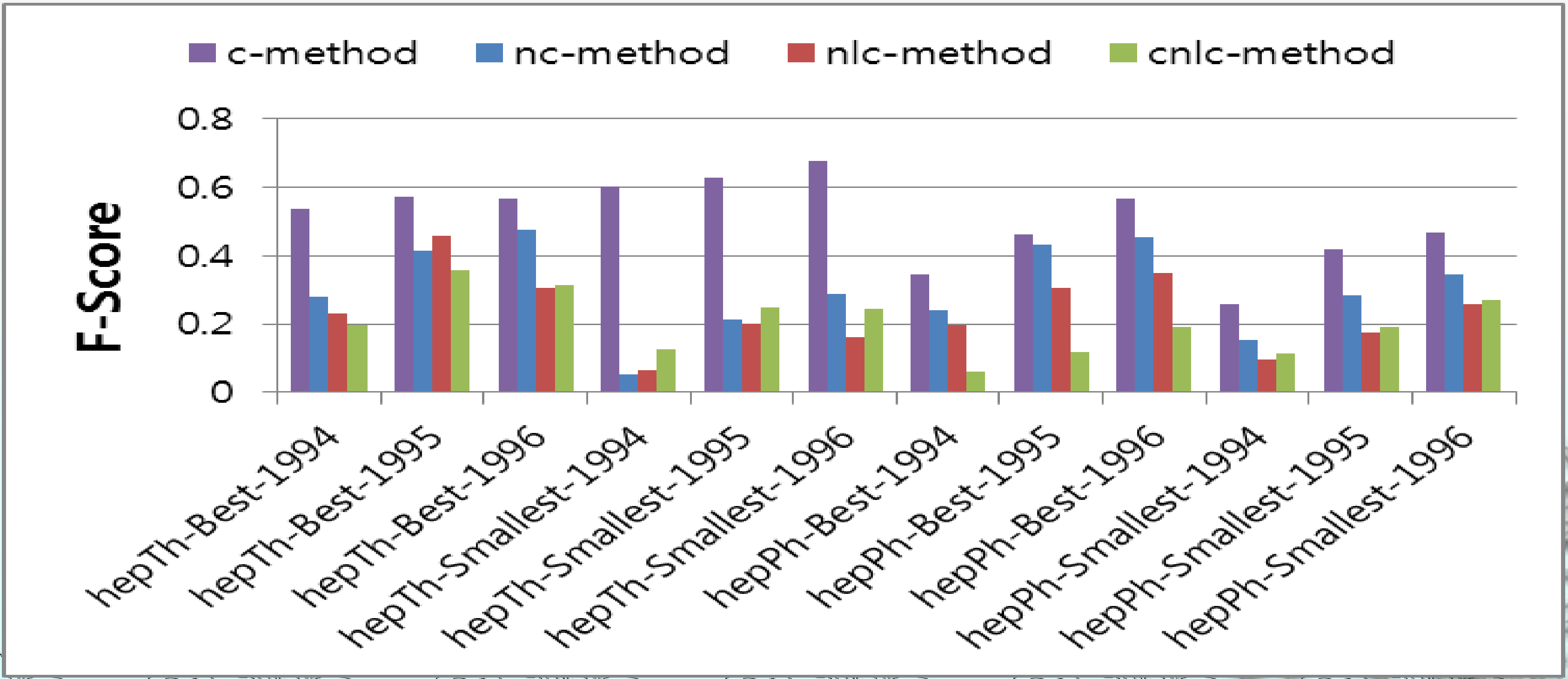
    Adding 1 – too little

    Adding 15 - overkill



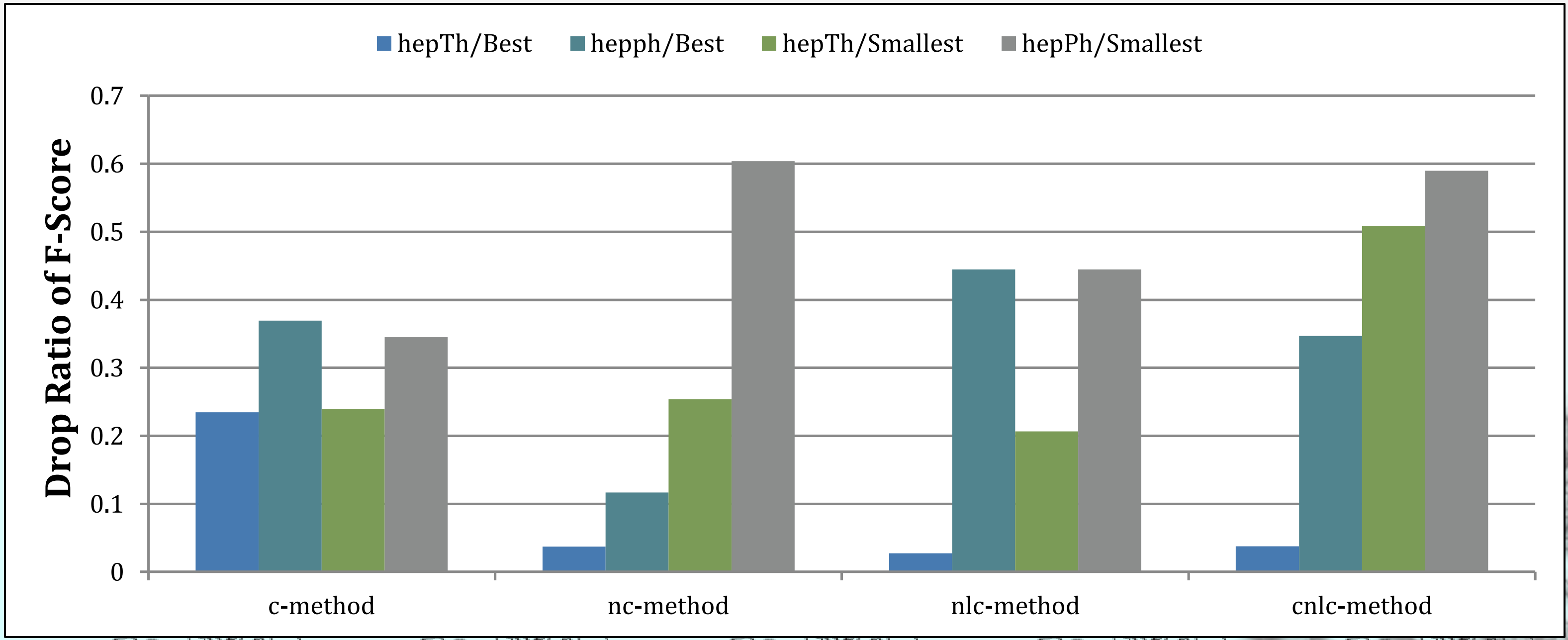
Results

- c-method outperforms other methods
  - Performance worsens as given graph becomes more dynamic
- Increase in data size increases performance



Results

- Repeating each method multiple times give predictions multiple years into the future
- Performance drop ratio is shown below
  - With large graph with fine-grained communities, nc-method is better than c-method.



Q/A