Indonesian Innovations on Information Technology 2013: Between Syntactic and Semantic Textual Network

Situngkir, Hokky

Bandung Fe Institute

17 September 2013

Indonesian Innovations on Information Technology 2013
Between Syntactic and Semantic Textual Network

Hokky Situngkir
[hs@compsoc.bandungfe.net]
Dept. Computational Sociology
Bandung Fe Institute

Abstract
Network and graph model is a good alternative to analyze huge collective textual data for the ability to reduce the dimensionality of the data. Texts can be seen as syntactic and semantic network among words and phrases seen as concepts. The model is implemented to observe the proposals of Indonesian innovators for implementation of information technology. From the analysis some interesting insights are outlined.

Keywords: innovation, semantic map, corpus, complex network, computational linguistics.

1. Introduction

Innovation is said not about ideas, but more about recognition of ideas [2]. Ideas are everywhere in the past and current way of collective thinking, but innovation makes the direction of how ideas evolve within the society, be it social, economic, and even political life [15]. Thus, seeing innovation is like the best way to hover around the current living dots of ideas and the way people innovate is connecting them one another by the celebration of new ideas, products, and even business and entrepreneurial platforms.

Along with some governmental institutions, there are some big corporations in Indonesia right now making a sort of "incubation" for new start-ups and young entrepreneurs in order to boost innovations in the country. They do the selections and the elected start-ups will be groomed in to the established business funded by investment. Before the selection phase, innovators around the country submit their proposal due to ideas, products, or even an ongoing business venture they were doing. In front of the jury for selection is hundreds of proposals to choose. Whatever the new innovators stated in their proposals are collectively the field of ideas of innovations among current Indonesian people. The proposals can be seen as a collection of corpus reflecting the need of innovative ideas among the society. Reading them one by one in details is one way to get the insight, but seeing it visually by statistically data crunching is one alternative way to get the big picture.

There are interesting patterns and properties by observing collection of texts as network and graph [14]. Graph representation can reduce the dimensionality of the text collections to gain insights instantaneously [5], thanks to the computational processing analysis. Network representation, widely known as semantic mapping, may reveal some interesting patterns represented by the corpus [3], not to mention that in some cases, gazing through the network visualizations may make people easier and faster to grasp interesting information within the corpus, rather than reading text by text.

To discuss about the information revealed within the large corpus of innovator's proposal is the main motivation of this paper. First, we discuss about the methodology review and acquisitions for textual analysis followed by general statistical properties of the collections of more than 300 proposals submitted by innovators in Indonesia to the National Telecommunication Company hosting the business incubation[*)]. Then the analysis delivers to the result and the discussions about the "face" of innovation related to information technology in Indonesia. The trends and major focus among Indonesian innovators brings the later discussions.

2. Network of Text

Conventionally, semantic graphs are built from the relations parsed from corpus. Computationally speaking, there is a database of relational semantic concepts or key phrases that is used to "read" the observed text or document [7]. However, the paper used an alternative different way to see textual documents. It is interesting to see that textual network representation can be seen in two perspective on corpus. We could build the graph

[*)] For more information about the program, refer to: `http:// www.indigoincubator.com`

representation either by seeing the syntactic relations of concept, words, or phrases and the semantic relations among concepts based on a relational database [10].

While semantic analysis tries to capture the semantic structure within sentences, the syntactic one is built by connecting words (and phrases) in sentences into an integrated whole. The latter concerns more about the emerged patterns among words and phrases while not necessarily the conceptual represented by the words (and phrases).

The idea of textual network representation presented in the paper is, nonetheless beyond the two distinctions. Words (and phrases) of which not grammatically sensitive (mostly nouns representing concepts) from each proposal document submitted for the refereeing process for the Incubation of Information Technology Innovation, are listed and modeled as fully connected graph of concepts.

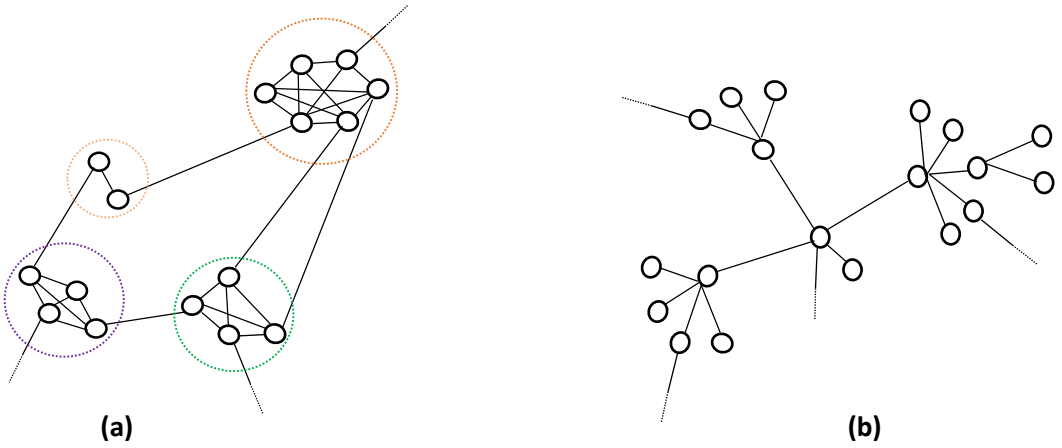

Figure 1. Syntactic (*a*) and semantic (*b*) network.

A word or a phrase is a node and the relation among nodes is the edge. We have a graph of $\Omega_w = (H_w, E_w)$ where $h_i \in H_w, (i = 1,2,3,...,N)$ as the member of set of N words/phrase and $E_w = \{e_{ij}\}$ as the set of the relations between word i and j when both are used within the same proposal. If two words/phrase/concept are used in a proposal, then the value of $e_{ij} = 1$, so that we have an undirected weighted graph with total weights are

$$< e_{ij} >= \sum_{i,j}^{N} e_{ij} \qquad (1)$$

and

$$< h_i >= \sum_{j}^{N} e_{ij} \qquad (2)$$

where $< e_{ij} >$ and $< h_i >$ denote the simultaneous use of words i and j and $< h_i >$ the word count of the word/phrase i. Somehow, the latter also represent the strength of the words i within the whole documents. From this representation we have the closed yet modular network represented by what innovators wrote in their proposals, due to the problems they faced, the products and ideas they suggest, with some fairly detail discussions on how the ideas and products can solve the stated problems.

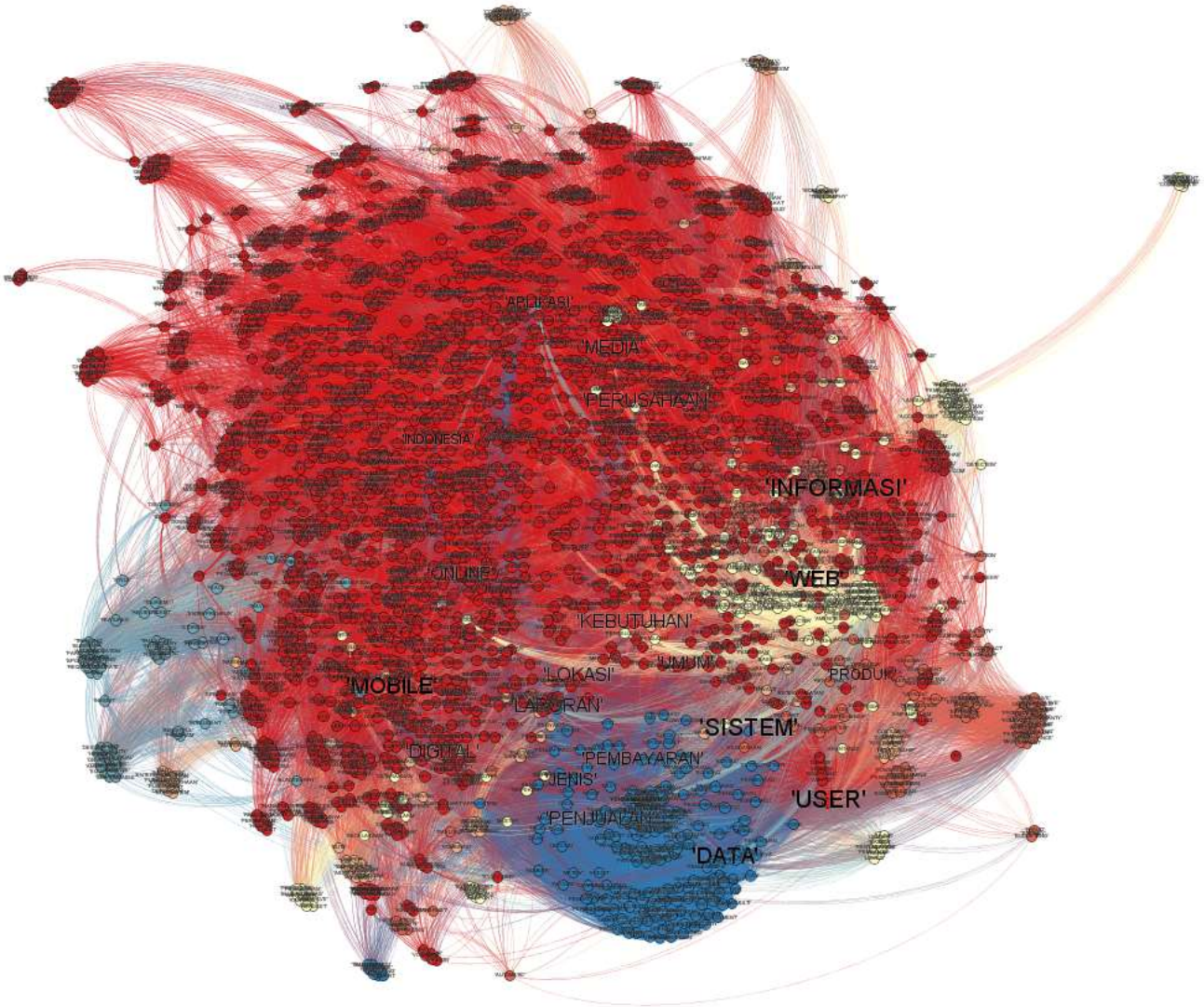

Figure 2. Syntactic network maps the total proposals
for innovations in information technology in Indonesia 2013.

From the yielded syntactic networks, we are also curious with the semantic-type network. We do this by reducing some edges to have a simpler image of the graph. To do this we use the minimum spanning tree algorithm [4]. This algorithm yield a mucho simpler network with tree representation: there is no closed loop within the network. Figure 1(b) illustrates the yielded semantic network.

The statistical properties of the textual network gives the signature of complexity, it has similarity with the social network. The semantic network is statistically sparse with small

amount of nodes have a relations with the vast majority of other nodes, has high local clustering, low average distances among nodes, and the power law degree distribution [16]. The specific statistical stylized facts of the network from Indonesian language is discussed in detail in [14] and is not main the motivation of this paper.

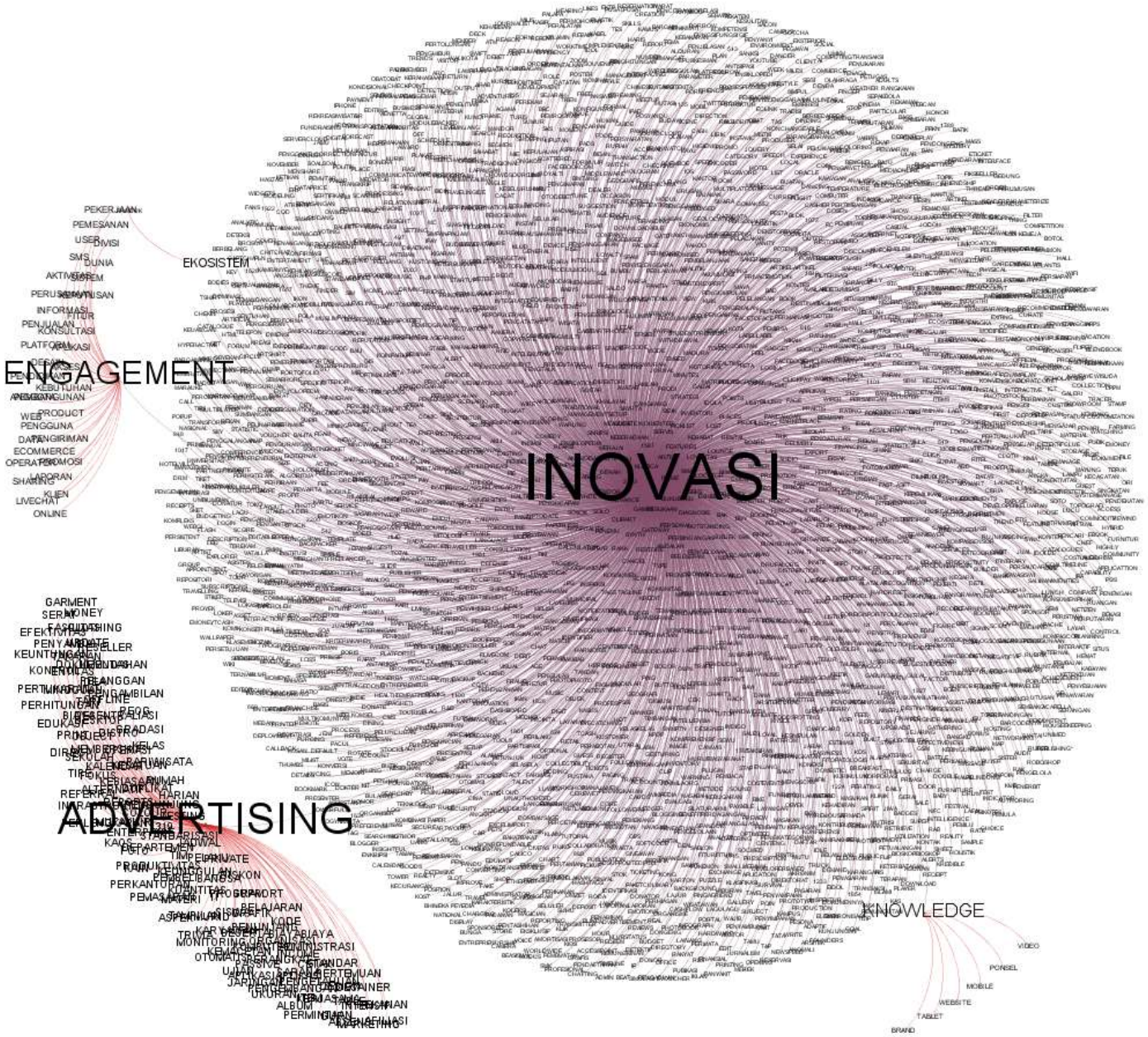

Figure 3. Semantic network maps the total proposals
for innovations in information technology in Indonesia 2013.

Table 1.
The comparison of 10th important concept of more than 300 proposals of Information Technology Innovations 2013 in Indonesia.

Syntactic Network	Semantic Network
PRODUK (*product*)	INOVASI (*inovations*)
APLIKASI (*applications*)	ADVERTISING
INDONESIA	ENGAGEMENT
USER	KNOWLEDGE
INFORMASI (*information*)	EKOSISTEM (*ecosystem*)
SISTEM (*system*)	LAYANAN (*services*)
MOBILE	CHARGE
DATA	GRATIS
WEB	CHARGING
FITUR (*features*)	EQUIPMENT

The main interest of the paper is to observe some important aspects related to the collective cognitive mapping of Indonesian innovators due to the vast submitted proposals of information technology innovations. We employ some centrality measures as discussed in more detail in previous work [13] to see important concepts within the network of the documents.

3. Reading by Drawing Indonesian Innovation Profile

The result of the graph drawing after reading the more than 300 proposals submitted for Information Technology Incubation for the year 2013 is shown in syntactic graph (in figure 2) and semantic graph (figure 3) via the minimum spanning algorithm. The graphs are made of 2687 conceptual words/phrases. It is a huge number of concepts, but in both figures we resize the label of the nodes proportional due to the respective degree within the network.

In advance, we were also curious about which concepts are central relative to one another in both of the graph representation. Be it the syntactic and semantic network, both are represented as adjacency matrix $A(i, j)$, with size $n \times n$ of n words/phrases in the graph. As discussed in detail in [13], we calculate the centrality measures by using the eigen vector of the *i*-th node as,

$$x(i) = \frac{1}{\lambda} \sum_{j:(j,i) \in V} x(j) \qquad (4)$$

where j denotes the node connected to i within the set of V and λ is a constant. In the terms of adjacency matrix, we can write,

$$x(i) = \frac{1}{\lambda} \sum_{j}^{n} A(i, j)\, x(j) \qquad (5)$$

and in the vector notation,

$$\vec{x} = \frac{1}{\lambda} \bar{A}\vec{x} \quad (6)$$

or

$$\bar{A}\vec{x} = \lambda\vec{x} \quad (7)$$

which is the standard eigen vector.
Ten most important words/phrases of each network are shown in table 1. Thus it is interesting to compare the centrality measures of both representations.

Table 2.
The 10th of the most important concepts from the big three domains of Information Technology Innovations 2013 in Indonesia.

	SYNTACTIC NETWORK	SEMANTIC NETWORK
INNOVATIONS ON PUBLIC SERVICE	FITUR (*feature*)	PEMBUKUAN (*booking*)
	PRODUK (*product*)	KESULITAN (*hardship*)
	APLIKASI (*application*)	JARINGAN (*network*)
	MEDIA	SAWIT (*palm*)
	USER	TANDAN (*cluster of leaves*)
	LAYANAN (*services*)	INFORMASI (*information*)
	INFORMASI (*information*)	PABRIK (*factory*)
	DATA	KELAPA (*coconut*)
	SOCIAL	PANEN (*harvest*)
	ONLINE	INTERNET_SYSTEM
INNOVATIONS ON EDUCATION	FITUR (*feature*)	ENTREPRENEURSHIP
	INDONESIA	JASA (*services*)
	ONLINE	TOKO (*stores*)
	INFORMASI (*information*)	STORE
	SISWA (*students*)	DIGITAL
	MATERI (*material*)	BOOK
	MEDIA	ENCYCLOPEDIA
	PLATFORM	ULASAN (*resume*)
	APLIKASI (*application*)	PORTAL
	SISTEM (*system*)	PEMBAJAKAN (*piracy*)
INNOVATIONS ON DIGITAL ENTERTAINMENT	MEDIA	LAYANAN (*services*)
	WEBSITE	GAME
	VIDEO	PERTARUNGAN (*fight, arcade*)
	FITUR	EDUKASI (*education*)
	SOCIAL	SERIAL (*series*)
	PRODUK (*product*)	PEMBAJAK (*pirate*)
	LAYANAN (*services*)	NEGARA (*country*)
	FACEBOOK	SKOR (*score, points*)
	INDONESIA	BADGE
	DUNIA (*world*)	JEJARING (*network*)

All of the submitted proposals are actually can be categorized within 9 basic classifications, namely, innovation in the theme of public service, education, digital media entertainment,

digital advertising, finance and banking, health applications, tourism applications, and transportation and logistics issues. However, there are only the first three categories are with highest number of applicants. Within the big three classifications, we do the similar calculations to see the most interesting and important issues as recognized and proposed by Indonesian innovators of information technology.

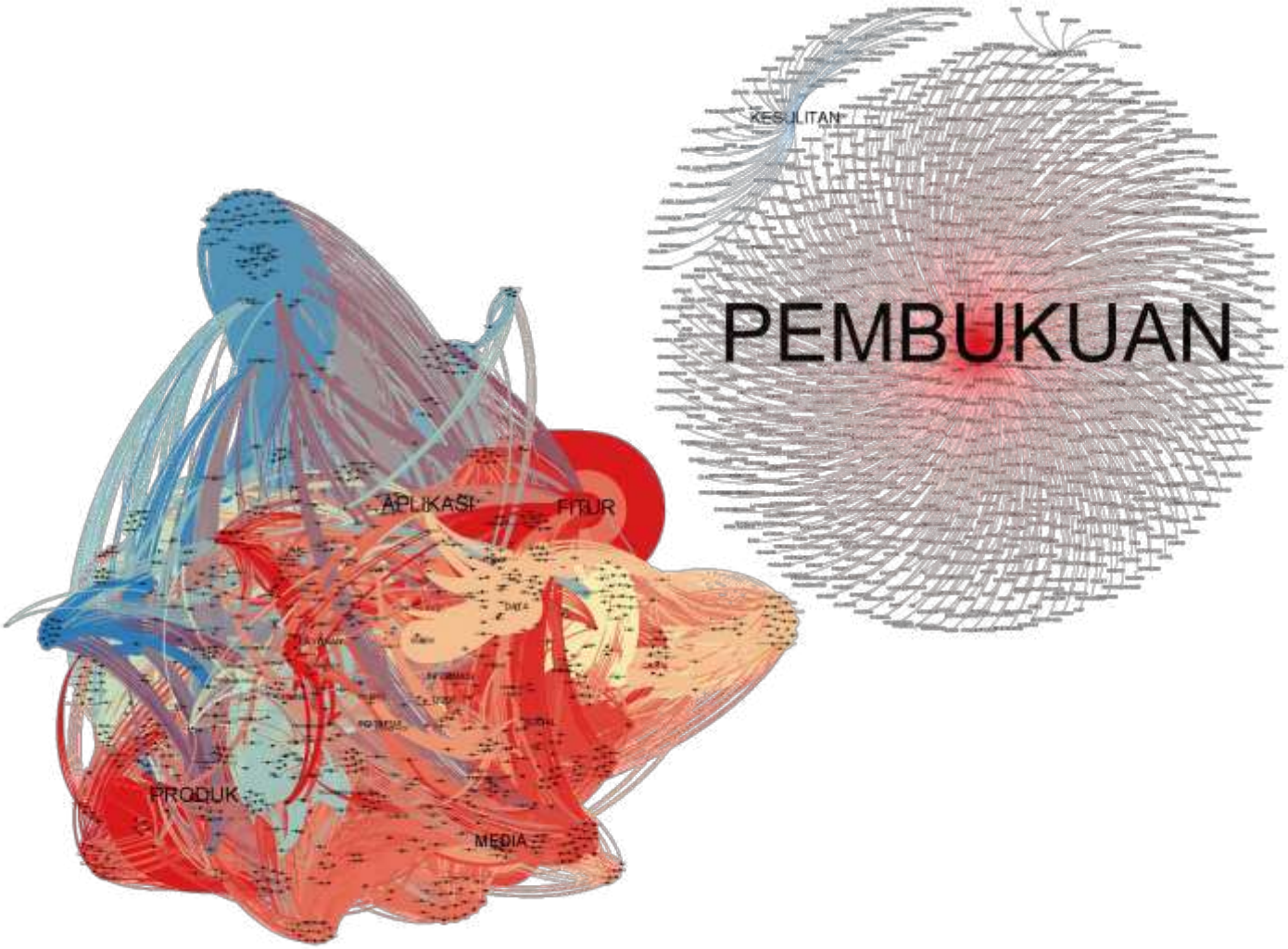

Figure 4. Syntactic (*left)* and the semantic network map the proposals for innovations in the topic of "public service".

4. Discussions

The result of the analysis has delivered the semantic mapping of the major concerns in the realm of Indonesian innovators when they face the issue of information technology. Syntactic map in figure 2 shows a more like "word cloud" depicting the relations among the concepts and how one proposal is related to one another. The words "system", "data" "mobile", "applications", "information", and "web" are the most used words within all of the submitted proposals. However, when it comes to the more hierarchical representation (figure 3), we can see that the word "engagement", "advertising", and "knowledge" are the main course of the whole proposals collectively. When we contrasted this finding with the centrality measures within the network, we can see that most innovators were talking about things related to the

products, and some technical issue due to growing and establishing the informational infrastructures (*e.g.:* data, applications, features). A little bit different with the one can be measured in the semantic network which is more about "how to develop business with the information technology" as shown by the importance of business-related words/phrases like "advertising", "cost", "service and goods", and so on.

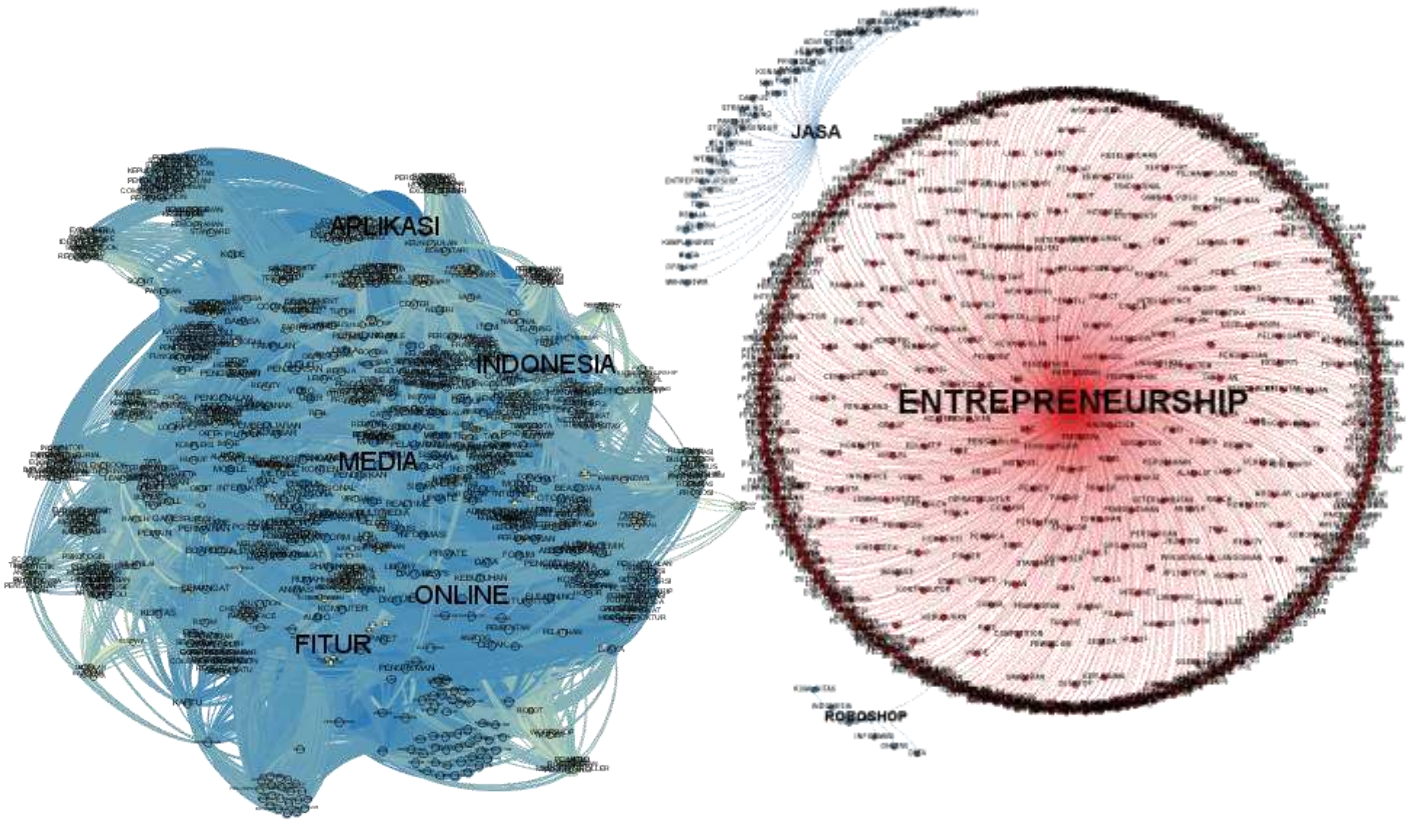

Figure 5. Syntactic (*left)* and the semantic network map the proposals for innovations in the topic of "education".

We can simply say that the attention of the innovators when they talk about the innovations in the information technology is actually about how to develop business even though they are using mostly some technical terminologies in their words and phrases. In a more detail observations, we can compare the global view with the one with more specific topics: public service, education, and digital entertainment.

When the innovators are talking about the acquisition of information technology into the public services, mostly it is about the aspects and features of the realm within the technology. They commonly use the terminologies like "media", "applications", "features", and so on. However, the semantic network reveals the concepts that are more "central" in the innovators' realm due to the public service issue are things like the employment of administration ("booking") in the public service domain. This also reflects the "hardship" among public trying to have access to public services, and how innovations in information technology propose help to resolve that. This is also confirmed in table 1.

An interesting property when we see the semantic mapping within the domain of education. Interestingly the innovators talked a lot about the importance of "entrepreneurship" while they have task to propose the ideas related to the acquisitions of information technology in to enhance educational system (figure 5). A more obvious issue is shown in table 2, where the innovators' proposal are actually speaking about some important issue related to "book publishing", "educational digital store", and one important aspect when we talk about information technology in the country: "piracy" due to the issue of violation of intellectual property right of software and digital products.

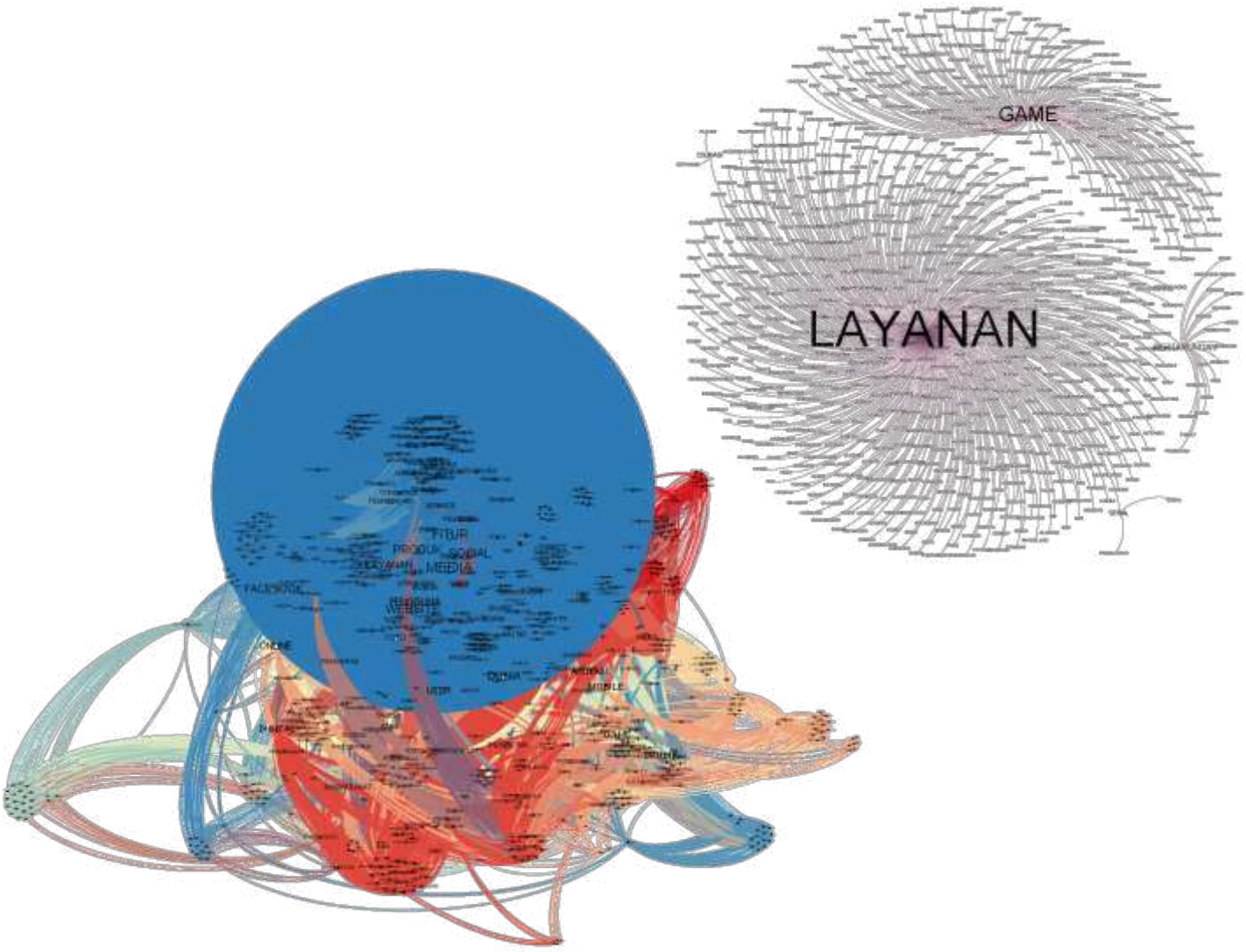

Figure 5. Syntactic (*left)* and the semantic network map the proposals for innovations in the topic of "digital entertainment".

When the innovations are about the digital entertainment, they concern more on the services ("*layanan*") and game computing. On the other hand, the digital entertainment is seen also related to the enhancement educational issue among users. Some products in the category of "digital entertainment" are also talking about education.

If we see the three sectors, the innovators talk more on similar issue in their proposals. They talk about information technology as media, features of applications, and business-related issue in the implementation of their ideas into services and goods. The minimum spanning tree reveals more detail and domain-specific aspects than the syntactic network.

Having these global and domain-specific properties of the syntactic and semantic networks of the innovative proposals for information technology, we capture the realms of the innovators and the aspects beneath the information technology-related terminologies. In further and deeper observations, we can see that actually we are now holding the possibility to see the "similarities" of ideas embedded within the proposals. This can be left for further implementation of the potential assistance of linguistic computation to read the huge amount of documents.

5. Concluding Remarks

Simplicity out of complexity is an important issue when we talk about the huge textual documents. The employment of semantic network to map concepts (words and phrases) used in the corpus may help to do this. The network representation can potentially be used to reduce the dimensionality of large amount of texts into some particular levels in which we can have instantaneous understanding of the global properties and stylized facts within corpus.

As we implement this into hundreds of proposals to for information-technology business incubation, we gain some important aspects due to the realms of Indonesian innovators in their endeavors for the acquisitions of information technology in specific domains, be it public services, education, entertainment, and so on.

Most proposals for the acquisitions of information-technology are about how to administer business development. The variations among topics they propose has dimmed the technical and specific aspects of the information technology they want to focus on. However, by using the analyses with the domain-specific semantic network, we can reveal some important and central theme they want to deliver in area of public services, education, and digital entertainment.

Acknowledgement

Author thanks Indra Purnama (*Telkom* Bandung Digital Valley) for discussions and data sharing. All faults remain author's.

Works Cited:

[1.] Barthelemy, M., Chow, E., & Eliassi-Rad, T. (2005). "Knowledge Representation Issues in Semantic Graphs for Relationship Detection". *Proceeding of AAAI Spring Symposium: AI Technologies for Homeland Security* 2005: 91-98.

[2.] Burkus, D. (2013). *The Myths of Creativity: The Truth About How Innovative Companies and People Generate Great Ideas*. Jossey-Bass.

[3.] Burk R., Chappell, A., Gregory, M., Joslyn, C., & McGrath, L. (2012). "Pattern Discovery Using Semantic Network Analysis." *Proceeding of Third International Workshop on Cognitive Information Processing (CIP)*, May 28-30, 2012, Baiona, Spain. IEEE.
[4.] Chen, C. & Morris, S. (2003). "Visualizing evolving networks: minimum spanning trees versus pathfinder networks". *Proceedings of the Ninth annual IEEE conference on Information visualization INFOVIS'03*: 67-74.
[5.] Corrêa, R. F. & Ludermir, T. B. (2007). "Dimensionality Reduction of very large document collections by Semantic Mapping". *Proceedings of 6th Int. Workshop on Self-Organizing Maps (WSOM) 08/2007*; 6(1).
[6.] Griffths, T. L. & Steyvers, M. (2002). "A probabilistic approach to semantic representation". *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.
[7.] Huang, C., Tian, Y. Zhou, Z., Ling, C. X., & Huang, T. (2006). "Keyphrase Extraction Using Semantic Networks Structure Analysis". *Proceeding of Sixth International Conference on Data Mining ICDM* 2006: 275-84.
[8.] Khanafiah, D. & Situngkir, H. (2006). Innovation as Evolution. Paper presented in the *5th Computation and Intelligence in Economics and Finance Conference*, Kaoh-Siung, Taiwan 2006.
[9.] Kim, D-H. & Jeong, H. (2003). "Scale-Free Spanning Trees of Complex Networks". *Journal of the Korean Physical Society* 44 (3): 624-7.
[10.] Liu, H. T. (2009). "Statistical properties of Chinese semantic networks". *Chinese Science Bulletin* 54: 2781—5.
[11.] O'Connor, K. (2003). *The Map of Innovation: Creating Something Out of Nothing*. Crown Business.
[12.] Situngkir, H. (2004). "On Selfish Memes: Culture as Complex Adaptive System". *Journal of Social Complexity* 2(1).
[13.] Situngkir, H. (2007). "Konsep Sentralitas dalam Jaringan Teks: Studi Editorial Media atas Politik Luar Negeri Indonesia terkait Isu Terorisme dan Kiprah Internasional Amerika Serikat. *BFI Working Paper Series* WPV2007. Bandung Fe Institute.
[14.] Situngkir, H. (2007). "Model Jaringan dalam Analisis Media". *BFI Working Paper Series* WPQ2007. Bandung Fe Institute.
[15.] Situngkir, H. (2009). "Evolutionary Economics Celebrates Innovation and Creativity-Based Economy". *The Icfai University Journal of Knowledge Management* 7(2):7-17.
[16.] Steyversa, M. & Tenenbaum, J. B. (2005). "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth". *Cognitive Science* 29: 41–78.