Results of plagiarism analysis from 2017-12-27 23:47 UTC

Date: 2017-12-27 15:20 UTC

All sources 45  Internet sources 30  Own documents 11  Organization archive 4

1. iopscience.iop.org/article/10.1088/1742-6596/801/1/012078/pdf
   60 matches

2. https://archive.org/stream/arxiv-1212.3906/1212.3906_djvu.txt
   40 matches

3. "CR-INT110-Semantic interpretation ...ot; dated 2017-10-09
   33 matches

4. iopscience.iop.org/article/10.1088/1742-6596/801/1/012022
   11 matches

5. iopscience.iop.org/article/10.1088/1742-6596/801/1/012020/meta
   11 matches

6. "CR-INT136-Social network extractio...ot; dated 2017-10-09
   12 matches

7. https://archive.org/stream/arxiv-1212.4702/1212.4702_djvu.txt
   10 matches

8. https://arxiv.org/pdf/1303.3964.pdf
   10 matches

9. eprints.binadarma.ac.id/id/2778/1/ICIBA2016-13-086-091-Nasution-SocialNetworkMining.pdf
   12 matches

10. https://arxiv.org/pdf/1604.06976.pdf
    12 matches

11. "CR-INT137-Enhancing to method for ...ot; dated 2017-10-09
    8 matches

12. "16. IOP.pdf" dated 2017-12-07
    7 matches

13. "17. IOP.pdf" dated 2017-12-07
    7 matches

14. "CR-INT135-Information Retrieval on...ot; dated 2017-10-09
    10 matches

15. "3028-3639-1-RV.pdf" dated 2017-10-30
    8 matches

16. "Text artikel no_11.pdf" dated 2017-12-06
    5 matches

17. www.academia.edu/3144214/Simple_Search_Engine_Model_Adaptive_Properties
    7 matches

18. "Abdullah_2017_J._Phys.__Conf._Ser._890_012102.pdf" dated 2017-10-12
    7 matches

19. https://www.researchgate.net/publication...ion_of_Search_Engine
    5 matches

20. https://www.researchgate.net/publication...and_Design_Patterns
    5 matches

21. www.academia.edu/3144197/Simple_Search_Engine_Model_Adaptive_Properties_for_Doubleton
    5 matches

22. https://link.springer.com/content/pdf/10.1007/978-3-319-05476-6_9.pdf
    6 matches

23. "31. IOP.pdf" dated 2017-12-07
    3 matches
5 documents with identical matches

[32] https://link.springer.com/chapter/10.1007/978-3-319-67621-0_20
0.9% 4 matches

[33] 2475-2791-1-RV.pdf dated 2017-09-26
0.7% 4 matches

[34] https://www.sciencedirect.com/science/article/pii/S2288430014500243
0.7% 1 matches

[35] https://vdocuments.site/chi-square-history.html
0.6% 2 matches

[36] citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.747.3201&rep=rep1&type=pdf
0.6% 2 matches

[37] 3472-4529-1-SM.pdf dated 2017-10-30
0.5% 4 matches

[38] https://www.researchgate.net/profile/Mah...n=publication_detail
0.6% 4 matches

[39] https://www.researchgate.net/profile/Sh...tion-perspective.pdf
0.4% 3 matches

[40] iopscience.iop.org/article/10.1088/1742-6596/822/1/012078/pdf
0.5% 2 matches

[41] iopscience.iop.org/article/10.1088/1757-899X/171/1/012078/pdf
0.5% 2 matches

[42] https://www.researchgate.net/publication...Adaptive_Properties
0.1% 2 matches

[43] www.finedictionary.com/Nasute.html
0.2% 2 matches

[44] 2668-3031-1-RV.pdf dated 2017-09-26
0.3% 2 matches

[45] https://rd.springer.com/content/pdf/10.1007/978-3-319-05476-6_9.pdf
0.3% 2 matches

[46] https://www.researchgate.net/publication...n=publication_detail
0.2% 1 matches

[47] https://research.google.com/pubs/AndreiBroder.html
0.2% 1 matches

[48] 3310-4262-1-RV.pdf dated 2017-10-30
0.2% 1 matches

[49] 2660-3022-1-RV.pdf dated 2017-09-26
0.2% 1 matches

[50] https://link.springer.com/chapter/10.1007/978-3-319-29544-2_7
0.1% 1 matches

[51] https://quizlet.com/128157095/formal-logic-3-flash-cards/
0.1% 1 matches

[52] https://www.gcflearnfree.org/internetbasics/using-search-engines/1/
0.1% 1 matches

9 pages, 5997 words

PlagLevel: selected / overall
101 matches from 53 sources, of which 30 are online sources.

Settings
Data policy: Compare with web sources, Check against my documents, Check against my documents in the organization repository, Check against organization repository, Check against the Plagiarism Prevention Pool
Sensitivity: Medium
Bibliography: Consider text
Citation detection: Reduce PlagLevel
Whitelist: --
Modelling and Simulation of Search Engine

To cite this article: Mahyuddin K M Nasution 2017 J. Phys.: Conf. Ser. 801 012078

View the article online for updates and enhancements.

Related content
- Studies on behaviour of information to extract the meaning behind the behaviour M K M Nasution, R Syah and M Elveny
- Improving PHENIX search with Solr, Nutch and Drupal
  - Dave Morrison and Irina Sourikova
- Locus Guard Pilot
  - Varsha Chandrashekar and Prabadevi B

Recent citations
- Mahyuddin K. M. Nasution et al
- Social Network Extraction Based on Web: A Comparison of Superficial Methods
  - Mahyuddin K. M. Nasution and Shahru Azman Noah
Modelling and Simulation of Search Engine

Information Technology, Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan
USU Medan 20155, Indonesia.

Abstract. The best tool currently used to access information is a search engine. Meanwhile, the
information space has its own behaviour. Systematically, an information space needs to be
familiarized with mathematics so easily we identify the characteristics associated with it. This
paper reveal some characteristics of search engine based on a model of document collection,
which are then estimated the impact on the feasibility of information. We reveal some of
characteristics of search engine on the lemma and theorem about singleton and doubleton, then
computes statistically characteristic as simulating the possibility of using search engine. In this
case, Google and Yahoo there are differences in the behaviour of both search engines, although
in theory based on the concept of documents collection.

1. Introduction
To access or search for information in an information space or system, we need tools [1]. One of tools
is the search engine, we know as a software system [2, 3]. In general, for helping to know and understand
a system, we use the model to assemble it such that mathematically a model can represent the search
engine [4]. Whereas, simulation can used for estimating the effect of search engine model on the
information space or system [5].

There are many different search engines. The search engine that arises naturally with the database or
search engine that grew up with the web (web search engine) [6, 7]. Dealing with the complexity of
information, the search engines helpless and disappear, the search engine shifts to meet the capabilities
required, or the search engines changed clothes and present be new. Therefore, all this will affect access
to information in space. In this case, the mathematical principle is not only used to systematize, but it
serves to optimize the creation of a search engine on information space. This paper aimed to express the
characteristics of search engine based on the constraints in the information space.

Suppose we denote the information space or system such as Ω [8]. The information space contain the
groups of documents or D [9]. Each group of documents consist of documents whereby there a word
w, i.e. the basic unit of discrete data, defined to be an item from a vocabulary indexed by \{1,\ldots,K\}, w_i
= 1 if k in K or w_i = 0 otherwise [10, 11]. Next, we define the terms related to the word.

Definition 1. A term \( t \) coincide with at least one or more words, i.e. \( \exists j \) \( \in \{1,\ldots,L\} \), \( k \neq \) \( j \), \( l \) is a number of parameters representing words \( w \), \( l \) is the number of vocabularies in \( t \) \( \# l \) is the size of \( n \).

Suppose that we have a term, that is a person name \( \# t = \{ \text{Mahyuddin Khairuddin Matyuso Nasution} \} \),
or \( \{w_1,w_2,w_3,w_4\} = \{\text{"Mahyuddin"","Khairuddin"","Matyuso"","Nasution"} \} \) as a set of words. We obtain
the power of sets $\{1, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\} = \{2^{|\Omega|}\}$. Note: $2^{|\{k\}|}$ is called $k$-th power of 2. In this case we have $|\{2^{|\Omega|}\}| = 15$ and $l = k$. Therefore, probability of $t$ or $p(t) = 1/(2^{|\Omega|})$. Suppose the vector space of $t$ is $\{2^{|\Omega|}\}$, we have a design for searching information in the search space by a software system as the search engine. We express it as follows [12, 13].

**Definition 2.** Suppose $\Omega$ is a set of documents indexed by search engine, i.e., a set consists of the ordered pair of the terms $t_o$ and documents $d_o$, or $(t_o, d_o), i=1, \ldots, l$. The relation table is two columns $t$ and $d$ as a representation of search engine whereby $\Omega = \{(t_o, d_o)\}$ is a subset of $\Omega$. The size of $\Omega$ is denoted by $|\Omega|$. 

**Definition 3.** Let $t$ be a search term and $q$ is a query, then $t \in q$ for $t$ in $d$, $d$ in $\Omega$.

In logical implication, Definition 3 express that a document is relevant to a query if it implies the query, that is if $d = q$ is true or $d = t$ is true for all $d$ in $\Omega$. $(d = t)$ is true. Thus, the degree of $d = q$ measured by $P(d = q)$. Therefore there are the uniform mass probability function for $\Omega$, i.e.

$$
P: \Omega \rightarrow [0,1]
$$

where $\Sigma P(d) = 1$.

**Definition 4.** Suppose $t$ is a search term or $t$ in $S$ whereby $S$ is a set of singleton search terms of search engine. A vector space $\Omega$, be a subset of $\Omega$, is a singleton search engine event (singleton) of documents that contain an occurrence of $t$ in $d_o$.

In the ordered degree of uncertainty of $d = t$ on $d = q$ means that

$$
P(\Omega_t) = P(\Omega_t | t = t) = \Sigma(d(\Omega, t = 1) | |\Omega | = |\Omega|).
$$

However, if search term in pattern, like $t = \{t_0, \ldots, t_n\}$, then a different result appears. In other words, $\Omega_p(\text{"x") = 1}$ if $t_i$ is true at $d$ in $\Omega$ exactly or $\Omega_p(\text{"x") = 0}$ otherwise, and the cardinality of $\Omega$ be $|\Omega| = \Sigma|\Omega_p(\text{"x")}| = 1$. In other word, each document that is indexed by search engine contains at least one occurrence about the search term. In degree of uncertainty of $d = t$ on $d = q$ is

$$
P(\Omega_p) = P(\Omega_p(\text{"x") = 1}) = \Sigma|\Omega_p(\text{"x")}| = |\Omega| |\Omega|.
$$

Thus $|\Omega_p| = |\Omega|$, so $|\Omega| \leq |\Omega|$, and $\Omega_p$ is a subset of $\Omega$.

Let $t$ and $t_r$ are two different search terms. If $t = t_r$, or $|t| = |t_r|$, then $\Omega_p$ be a subset of $\Omega$ or $\Omega_r$, be a subset of $\Omega$, or $\Omega_r$ be a subset of $\Omega$.

Let $t$ and $t_r$ are search terms, refer to the definitions above, will be revealed some characteristics related to the search engine as a system. All characteristics derived from the adaptation formula that build model of the problem completion relating to the possible results of the search engine. Some of the adaptive characteristics are as follows [12, 13, 14].

**Lemma 1.** If $t \neq t_r$, and $t \cap t_r = \emptyset$, then $|\Omega \cap \Omega_t | = 0$ and $|\Omega \cup \Omega_r | = |\Omega| = |\Omega|$. Here $\Omega$ and $\Omega_t$ are subsets of $\Omega$.

**Proof.** If $t \neq t_r$, and $t \cap t_r = \emptyset$ mean that for all $w$ in $t$, all $w$ not in $t_r$ and all $w$ in $t_r$ all $w$ not in $t$, then for all $w$ in $d$, all $w$ not in $d$, and for all $w$ in $d_r$ all $w$ not in $d_r$ such that $t \cup t_r = t \cup t_r$ and $d \cup d_r = d \cup d_r$. 

1. Mahyuddin Khairuddin Matuyso Nasution.
Therefore, \( \Omega = \{(t, \varnothing)\} \) and \( \Omega = \{(t, d)\} \) are two independent events from queries, or \( t \) and \( t_0 \) are true at \( d \) in \( \Omega \), respectively. In this case, \( \Omega \cap \Omega = \varnothing. \) In other words, \( \{(t, d)\} \cup \{(t, d)\} = \Omega \cup \Omega. \) Therefore, we have

\[
|\Omega \cap \Omega| = 0
\] (3)

and

\[
|\Omega \cup \Omega| = |\Omega| + |\Omega| + \ldots + |\Omega|
\] (4)

Lemma 2. \( \forall t \neq t_0 \neq \varnothing \) and \( |t| \neq t_0, \) then \( |\Omega| = |\Omega| \) where \( \Omega \) and \( \Omega \) are subsets of \( \Omega. \)

Proof. Based on assumption, we have for all \( w \) in \( t \), all \( w \) in \( t_0 \), but there are \( w \) in \( t \) whereby \( w \) not in \( t_0 \) such that \( t \cap t = \varnothing \) and \( t \cup t = \varnothing. \) Similar concept, for all \( w \) in \( t \), all \( w \) in \( t_0 \), and because all \( w \) in \( t \), we conclude that \( w \) also in \( d \), but there are \( w \) in \( t \) and \( w \) in \( d \), whereby not in \( t_0 \) such that \( w \) not in \( d \). Thus, if \( t \neq t_0 \), then \( t \cap t = \varnothing \) and if \( t \neq t_0 \) then \( t \cap t = \varnothing. \) Therefore, \( \Omega = \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \Omega \cup \Omega. \) In other words,

\[
|\Omega| = |\Omega| + |\Omega| + \ldots + |\Omega|
\] (5)

Proposition 1. \( \forall t \neq \ldots \neq t \neq t_0 \) and \( |t| \neq t_0 \), then \( |\Omega| = |\Omega| \) where \( \Omega \) and \( \Omega \) are subsets of \( \Omega. \)

Proof. Based on generalization of Equation (3) and Equation (4), we derive \( |\Omega| = |\Omega| + |\Omega| + |\Omega| + \ldots + |\Omega| \) where \( \Omega \), \( \ldots \), \( \Omega \), are subsets of \( \Omega. \)

Lemma 3. \( \forall t \neq t_0 \neq \varnothing \) and \( \Omega \cap \Omega \neq \varnothing \), then \( |\Omega| = |\Omega| \) are subsets of \( \Omega. \)

Proof. \( t \neq t_0 \neq \varnothing \) and \( \Omega \cap \Omega \neq \varnothing \) mean that for all \( w \) in \( t \), all \( w \) not in \( t_0 \), and for all \( w \) in \( t \), all \( w \) in \( t_0 \), then \( t \cap t_0 = \varnothing \) but for all \( w \) in \( d \), there are \( w \) in \( d \), and for all \( w \) in \( d \), there are \( w \) in \( d \) also, then \( d \cap d_0 = d \cap d_0 \) where \( \Omega \cap \Omega \neq \varnothing \) and \( \Omega \cap \Omega \neq \varnothing \) we obtain \( \Omega \cap \Omega = \{(t, d)\} \cap \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \Omega \cup \Omega \) or \( \Omega \cap \Omega = \Omega \cap \Omega \) or

\[
\Omega \cap \Omega = \Omega.
\] (6)

Similarly,

\[
\Omega \cup \Omega = \Omega.
\] (7)

In other words, \( \Omega = \{(t, d)\} \cup \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \{(t, d)\} \cup \{(t, d)\} = \Omega \cup \Omega \) and

\[
|\Omega \cup \Omega| = |\Omega| + |\Omega| + \ldots + |\Omega|
\] (8)

Definition 3. \( \forall t \neq t_0 \neq \varnothing \) let \( t \neq t_0 \) and \( t \) in \( S \), where \( S \) is a set of singleton search term of search engine. \( A \) is a doubleton search term in \( S \) whereby the vector space of doubleton search term denoted by \( \Omega \cap \Omega \) is a doubleton search engine event of documents that contain a co-occurrence of \( t \) and \( t_0 \) such that \( t \neq t_0 \) and \( t \neq t_0 \) in \( d \) whereby \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \) are subsets of \( \Omega. \)

Theorem 1. \( \forall t \neq t_0 \neq \varnothing \) let \( t \neq t_0 \) and \( t \) in \( S \), where \( S \) is a set of singleton search term of search engine. \( A \) is a doubleton search term in \( S \) whereby the vector space of doubleton search term denoted by \( \Omega \cap \Omega \) is a doubleton search engine event of documents that contain a co-occurrence of \( t \) and \( t_0 \) such that \( t \neq t_0 \) and \( t \neq t_0 \) in \( d \) whereby \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \) are subsets of \( \Omega. \)

Therefore, \( \Omega \cap \Omega \) is a doubleton search engine event of documents that contain a co-occurrence of \( t \) and \( t_0 \) such that \( t \neq t_0 \) and \( t_0 \) whereby \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \) are subsets of \( \Omega, \) \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \), \( \Omega \).
also are random.

in the list of singletons respectively to test of sample. The next columns contain the list of singletons respective to each other. In other words, \( \{t,s\} = \{(t,1),(t,2),(t,3)\} \). Therefore, \( |\Omega \cup \Omega_2| \leq |\Omega| \) and \( |\Omega \cap \Omega_2| \leq |\Omega| \) for all search terms \( t \) and \( t_s \).

**Corollary 1.** If \( t \) and \( t_s \) are the different search terms, then \( |\Omega \cup \Omega_2| = |\Omega \cap \Omega_2| + |\Omega \cap \Omega_2| \).

**Proof.** As the direct or indirect consequence of Proposition 1 and Theorem 1.

The purpose of simulation, in this case, is to construct an approach for selecting the documents in information space or for disclosing the information in the repository [15]. As an experiment to collect data, which is to select \( n \) objects from the community. For example, we collect data from the academic community of Faculty of Medicine University of Sumatera Utara (USU), i.e. \( n = 51 \) academic actors, or in a list is \( A = \{Abdul Majid, Abdul Rachman Saragih, Abdul Rasyid, Abdullah Affi Siregar, Achanuuddin Hanafie, Adi Kunuma Aman, Alfred C. Satyo, Azkarolleah Aboet, Atan Baas Simuahji, Ayodhya Pitaloka Pasaribu, Aznan Lelo, Bachtiar Surya, Budi R. Hadibroto, Chairuddin Panusuman Lubis, Chairul Yoei, Darwin Dallunthe, Daulat Hasiholan Sibuea, Delfi Lutan, Delfitri Munir, Erwin Dharma Kadar \}. Among the names of actors as the term, two different terms \( t \) and \( t_s \) have several options that can correspond to words of each name, such as mutual, including, or intersection. Therefore, each term has the opportunity to be placed in the position of a particular index. The position of each term in the search engines for example based on the selected collection of a number of documents related to the term.

| Actor Name (A) | Medium of randomness test | Search engine as test simulation | Search engine as comparative simulation |
|----------------|----------------------------|---------------------------------|----------------------------------------|
|                | \( |\Omega| \)         | \( |\Omega\cup\Omega_1| \) | \( |\Omega\cap\Omega_1| \) | \( |\Omega| \) | \( |\Omega\cup\Omega_2| \) | \( |\Omega\cap\Omega_2| \) |
| a              | 0 or 1                     | 0 or 1                          | 0 or 1                                 | 0 or 1 | 0 or 1                                 | 0 or 1 |
| ...            | ...                        | ...                             | ...                                    | ...    | ...                                    | ...    |
| b              | 0 or 1                     | 0 or 1                          | 0 or 1                                 | 0 or 1 | 0 or 1                                 | 0 or 1 |
| ...            | ...                        | ...                             | ...                                    | ...    | ...                                    | ...    |
| ...            | ...                        | ...                             | ...                                    | ...    | ...                                    | ...    |
| z              | 0 or 1                     | 0 or 1                          | 0 or 1                                 | 0 or 1 | 0 or 1                                 | 0 or 1 |
| Average        | pr or lc                   | a or n                          | avg                                    | avg_2  | avg                                    | avg_4  | avg_6  |
| \( n_1 \)      | \( n_1 \)                  | \( n_1 \)                       | \( n_1 \)                              | \( n_1 \) | \( n_1 \)                              | \( n_1 \) |
| \( n_2 \)      | \( n_2 \)                  | \( n_2 \)                       | \( n_2 \)                              | \( n_2 \) | \( n_2 \)                              | \( n_2 \) |
| \( \mu_r \)    | \( \mu_r \)                | \( \mu_r \)                     | \( \mu_r \)                            | \( \mu_r \) | \( \mu_r \)                            | \( \mu_r \) |
| \( \sigma_r \) | \( \sigma_r \)             | \( \sigma_r \)                  | \( \sigma_r \)                          | \( \sigma_r \) | \( \sigma_r \)                          | \( \sigma_r \) |

In the sample that can represent population, we develop a table of information as experiment design for providing data, Table 1. Data that reveal characteristics of a search engine. In the table, the first column is the actor’s names alphabetically ordered. The second column contains academic level: It is used to test whether the sample is random, the academic level as medium of randomness test (mrt). The third column involves data of scientific publications indexed by Scopus whereby the actor consists of two categories: the author or not, data of scientific publications as the comparative mrt. It is intended to support the randomness test of sample. The next column contains the list of singletons respective to \( t \) and \( t_s \) in quotes, and a list of doubletons of \( t \) and \( t_s \) (singleton with keyword). In this case, we ensure that the singletons also are random.
In general, the information space consisting of documents viewed as the population. Statistically, the population is random, and it was tested whether the characteristics also lowered to the sample, so that any measurement about sample describe population. We separate the sample into two categories: number of first categories

\[ n_1 = \sum_{i=1}^{n_1} a_{i1} \]  
\[ n_2 = \sum_{i=1}^{n_2} a_{i2} \]

whereby \( a_i \) is elements of \( A \) that meet first category and \( a \) is elements of \( A \) that meet second category. While \( r \) is how many times the category change in the sample. Thus, the average of \( r \) is

\[ \mu_r = \frac{2(n_1n_2(n_1+n_2))}{n_1+n_2} + 1 \]

and the variance of \( r \) is

\[ \sigma_r^2 = \frac{(2n_1n_2(2n_1n_2-n_1n_2))((n_1+n_2)^2(n_1+n_2-1))}{n_1+n_2} \]

Then, we have \( Z_{count} \) as follows

\[ Z_{count} = \frac{r - \mu_r}{\sigma_r} \]

for hypotheses used are as follows: \( H_0 \): the data sequence is random, and \( H_1 \): the data sequence is not random. For academic level as category: professor (pr) or lecturer (lc), we have \( n_p = 34 \) and \( n_l = 17 \). By using Equations (3), (4), and (5), we obtain \( \mu_r = 23.67, \sigma_r = 9.93 \) and \( Z_{count} = -1.79 \), and for \( a = 0.05 \) we obtain \( Z_{a=0.025} = 1.96 \leq Z_{count} \leq Z_{a=0.025} = 1.96 \), and because \( r \) is located between the critical value then the decision is received \( H_1 \). Seen from the publication of scientific papers indexed by Scopus: author (a) or not (n), we have the similar conditions such that the sequence of data is random.

Furthermore, to test the randomness perfectly, tested independence of two data space by using chi-square \((\chi^2)\). Suppose the data space \((dx)\) is presented in matrix form as follows,

\[
\begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1n} \\
  x_{21} & x_{22} & \cdots & x_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nn}
\end{pmatrix}
\]

Amount of data \( x_j \) is \( S_j \) as follows

\[ S_j = \sum_{i=1}^{n} x_{ij} \]

So that we can calculate the expectations of each data as follows

\[ e_{11} = \frac{\left( \sum_{i=1}^{n} x_{i1} \right) \left( \sum_{j=1}^{n} x_{j1} \right)}{S_{11}} \]
\[ e_{12} = \frac{\left( \sum_{i=1}^{n} x_{i1} \right) \left( \sum_{j=1}^{n} x_{j2} \right)}{S_{11}} \]
\[ \vdots \]
\[ e_{nn} = \frac{\left( \sum_{i=1}^{n} x_{i1} \right) \left( \sum_{j=1}^{n} x_{jn} \right)}{S_{11}} \]

and we have a matrix of expectations as follows

\[
\begin{pmatrix}
  e_{11} & e_{12} & \cdots & e_{1n} \\
  e_{21} & e_{22} & \cdots & e_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  e_{n1} & e_{n2} & \cdots & e_{nn}
\end{pmatrix}
\]
Amount of data $d_j$ is $E_j$ as follows

$$E_j = \sum_{i=1,...,n, j=1,...,m} c_{ij}$$  \hspace{1cm} (8)

Then, we have

$$\chi^2 = \sum_{i=1,...,n, j=1,...,m} (x_{ij} - e_{ij})^2 / e_{ij}$$  \hspace{1cm} (9)

with degree of freedom $(df)$ is $(m-1)(n-1)$. For example, among 51 actor names we have $n = 4$ professors, $x_1 = 17$ lectures, $x_2 = 17$ authors, and $x_3 = 34$ non-authors. Based on Eq. (7) we can calculate their expectations, i.e. $e_1 = e_{12} = e_{21} = e_{22} = 25.5$, and based on Eq. (9) we obtain $\chi^2 = 21.97$ for test statistic $T$ as chi-squared distribution with $(m-1)(n-1) = (2-1)(2-1) = 1$ degree of freedom and the acceptance region for $T$ with a significance level of 5% is $\chi^2_{0.05} = 3.841$. This tell us there is a relationship between type of academic level and authors.

To reveal characteristics of search engine based on a model, we conduct an experiment about singleton and doubleton of Google search engine as test simulation and of Yahoo search engine as comparative simulation as follows.

1. Randomness test: Calculate the randomness test for $t_r$, $t_s$, and $t_{r,s}$ by completing the computations as follows:
   a. For $t_r$, we have the amount of 51 $|\Omega|$ from Google search engine, that is 2635514 with average $(\text{avg}) = 51676.75$. Number of $|\Omega|$ greater than or equal to $\text{avg}$ is 41, while number of $|\Omega|$ less than $\text{avg}$ is 10. By using Eq. (3), Eq. (4), and Eq. (5), $Z_{\text{avg}} = 6.54 \approx -0.025 = -1.96$. Therefore, reject $H_0$ and 51 singletons of $t_r$ from Google search engine is not random.
   b. However, the amount of 51 $|\Omega|_r$ from Google search engine, that is 1095045 with average $(\text{avg}) = 21471.47$. Number of $|\Omega|_r$ greater than or equal to $\text{avg}$ is 3, $|\Omega|_r$ less than $\text{avg}$ is 48. By using the similar equations, $Z_{\text{avg}} = -1.50 \approx -0.025 = -1.96$, and $H_0$ accepted whereby 51 singletons of $t_r$ from Google search engine is random.
   c. For doubleton $t_{r,s}$, whereby $t_r$ = “Universitas Sumatera Utara” as a keyword, we have amount of 51 $|\Omega_1 \setminus \Omega_2|$ from Google search engine, i.e 61092 with $(\text{avg}) = 1197.88$. Number of $|\Omega_1 \setminus \Omega_2|$ greater than or equal to $\text{avg}$ is 15 and number of $|\Omega_1 \setminus \Omega_2|$ less than $\text{avg}$ is 36. With that, we obtain $Z_{\text{avg}} = -1.31 \approx -0.025 = -1.96$ based on Eq. (3), Eq. (4) and Eq. (5), and $H_0$ accepted, thus 51 doubletons of $t_{r,s}$ from Google search engine is random.
   d. Whereas, for $t_r$, by using Yahoo search engine, we have the amount of 51 $|\Omega|$ is 2365061 with $(\text{avg}) = 46373.76$. So $n(pr) = 5$ and $n(lc) = 46$. $Z_{\text{avg}} = -0.03 = -1.96$. On that basis, $H_0$ accepted, thus 51 singletons of $t_r$ from Yahoo search engine is random.
   e. Similarly for $t_s$, the amount of 51 $|\Omega_s|_r$ from Yahoo search engine, that is 395815 with $(\text{avg}) = 7766.08$. Number of $|\Omega_s|_r$ greater than or equal to $\text{avg}$ is 3, $|\Omega_s|_r$ less than $\text{avg}$ is 48. By using the similar equations, $Z_{\text{avg}} = -1.50 \approx -0.025 = -1.96$, and $H_0$ accepted whereby 51 singletons of $t_s$ from Yahoo search engine is random.
   f. For doubleton $t_{r,s}$, whereby $t_r$ = “Universitas Sumatera Utara” as a keyword, we have amount of 51 $|\Omega_1 \setminus \Omega_2|$ from Yahoo search engine, i.e 15361 with $(\text{avg}) = 301.19$. Number of $|\Omega_1 \setminus \Omega_2|$ greater than or equal to $\text{avg}$ is 12 and number of $|\Omega_1 \setminus \Omega_2|$ less than $\text{avg}$ is 39. With that, we obtain $Z_{\text{avg}} = -0.42 \approx -0.025 = -1.96$ based on Eq. (3), Eq. (4) and Eq. (5), and $H_0$ accepted, thus 51 doubletons of $t_{r,s}$ from Yahoo search engine is random.
2. Independence test: For a contingency table has $m$ rows and $n$ columns, a test of independency that null and alternative hypotheses are:

$H_0$: The two or more categorical variables are independent.  
$H_1$: The two or more categorical variables are related.

| Table 2. Samples and categories | Google search engine | Yahoo search engine |
|--------------------------------|----------------------|--------------------|
| Categories                     | $|\Omega|$ | $|\Omega_{v}|$ | $|\Omega| \cap |\Omega_{v}|$ | $|\Omega|$ | $|\Omega_{v}|$ | $|\Omega| \cap |\Omega_{v}|$ |
| $n_1$                          | 10 | 3 | 15 | 5 | 3 | 12 |
| $n_2$                          | 41 | 48 | 36 | 46 | 48 | 39 |

a. First, we test the independence $|\Omega|$ of Google and $|\Omega|$ of Yahoo. By using Eq. (6), Eq. (7), Eq. (8), and Eq. (9) toward $n(|\Omega| \cap |\Omega_{v}|)$ see Table 2, we obtain $\chi^2 = 1.95 3.84$ with $df = 1$, and $H_0$ accepted for $\alpha = 0.05$. Thus two samples are independent.

b. Second, we test the independence $|\Omega_{v}|$ of Google and $|\Omega_{v}|$ of Yahoo. By using similar equations against $n(|\Omega_{v}|)$ see Table 2, we have obtain $\chi^2 = 0.00 3.84$ with $df = 1$, and $H_0$ accepted for $\alpha = 0.05$. Thus two samples are independent.

c. Third, we test the independence $|\Omega| \cap |\Omega_{v}|$ of Google and $|\Omega| \cap |\Omega_{v}|$ of Yahoo. By using similar equations with $n(|\Omega| \cap |\Omega_{v}|)$ see Table 2, we get value of $\chi^2 = 0.45 3.84$ with $df = 1$, and $H_0$ accepted for $\alpha = 0.05$. Therefore, two samples are independent.

d. For getting behavior of $|\Omega| \cap |\Omega_{v}|$, we test independence among singletons and doubleton of Google search engine. By using Eq. (6), Eq. (7), Eq. (8), and Eq. (9) for $n(|\Omega| \cap |\Omega_{v}|)$ see Table 2, we obtain $\chi^2 = .53 .82$ and $df = 3$, and $H_0$ rejected for $\alpha = 0.05$. Therefore, three samples of Google search engine are dependent.

e. In contrast to that, we test independence among singletons and doubleton of Yahoo search engine. Based on similar concept, we obtain $\chi^2 = 7.71 7.82$ with $df = 3$, and $H_0$ accepted for $\alpha = 0.05$. Therefore, three samples of Yahoo search engine are independent.

f. Therefore, for all characteristics in Table, based on Eq. (6), Eq. (7), Eq. (8), and Eq. (9), the $\chi^2 = 18.98$ greater than 12.59 for $df = 6$ and $\alpha = 0.05$ such that $H_0$ rejected. Therefore, all the data as a whole is dependent.

In general, a collection of documents in information space and indexed by a system be random, see randomness test (1a, 1c, 1d, 1e, and 1f), and information space $\Omega$ has a normal distribution, where Eq. (9) be the **uniform mass probability function**. A row of data in $A$ is random with a confidence level of 95%.

Although the same characters can be derived based on set theory, but singleton from different search engines are not interdependent. So the information presented freely with each other, caused by each search engine has its own potential and capabilities. There are different potential between Google search engine and Yahoo search engine. In Google search engine, the singletons and doubletons are dependent. Whereas in Yahoo search engine, the singleton and doubleton are independent. Therefore, an information space such as system have information tied to each other, but in different sub-systems can be built mutually bound: Google search engine and Yahoo search engine, for example, as different subsystems.
To model and simulate the search engines has been developed the adaptive and selective approach. Adaptive approach produced some formal characteristics while the selective approach generates the characteristic in reality. Both reveal the possibility of the differences about the information presented by the search engine although they has same basic concept. For example, the Google and Yahoo search engines show the different behavior. Further research will reveal some other formulation and characteristic of search engine.

References

[1] Fredrik K, Andersson, and S D Silvestrov 2008 The mathematics of internet search engines Acta Appl Math 104.

[2] M Haman, K Lakhotia, J Singer, D. R. White, and S Yoo 2013 Cloud engineering is search based software engineering too The Journal of Systems and Software 86.

[3] R Roj 2014 A comparison of three design tree based search algorithms for the detection of engineering parts constructed with CATIA V5 in large databases Journal of Computational Design and Engineering 1(3).

[4] A Anagnostopoulos, A Broder, and K Punera 2008 Effective and efficient classification on a search-engine model Knowl Inf Syst 16.

[5] G Meghabghab and A Kandel 2004 Stochastic simulations of web search engines: RBF versus second-order regression models Information Sciences 159.

[6] M Song, I-Y Song and P P Chen 2004 Design and development of a cross search engine for multiple heterogeneous database using UML and design patterns Information System Frontiers 6.

[7] S Agrawal, K Chakrabarti, S Chaudhuri, V Ganti, A C Konig, and D Xin 2009 Exploiting web search engines to search structured databases WWW, Madrid, Spain, ACM.

[8] M Tvarozek and M Bielikova 2007 Adaptive faceted browser for navigation in open information spaces WWW Banff, Alberta, Canada ACM.

[9] J G Davis, E Subrahmanian, S Kanda, H Granger, M Collins and A W Westerberg 2001 Creating shared information spaces to support collaborative design work Information Systems Frontiers 3(3).

[10] M K M Nasution and S A Noah 2012 Information retrieval model: A social network extraction perspective IEEE International Conference on Information Retrieval & Knowledge Management.

[11] M K M Nasution 2014 New method for extracting keyword for the social actor Intelligent Information and Database Systems LNAI 8397.

[12] M K M Nasution 2012 Simple search engine model: Adaptive properties Cornell University Library.

[13] M K M Nasution 2012 Simple search engine model: Selective properties Cornell University Library.

[14] M K M Nasution 2011 Kolmogorov complexity Clustering and similarity Bulletin of Mathematics 3(1).

[15] M K M Nasution 2013 Simple search engine model: Selective properties Cornell University Library.