IOP Conf. Series: Journal of Physics: Conf. Series **801** (2017) 012078 doi:10.1088/1742-6596/801/1/012078

Modelling and Simulation of Search Engine

Mahyuddin K M Nasution[1*]

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

[*]Email: mahyuddin@usu.ac.id

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.

2. Basic Concept and Motivation

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 d_j whereby there a word w, i.e. the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1,\ldots,K\}$, $w_k = 1$ if k in K or $w_k = 0$ otherwise [10, 11]. Next, we define the terms related to the word.

Definition 1. A term t_x coincide with at least one or more words, i.e. $t_x = (w_l|l=1,\ldots,L)$, $k \leq l$, l is a number of parameters representing words w, l is the number of vocabularies in t_x, $|t_x| = l$ is the size of t_x.

Suppose that we have a term, that is a person name t_l = 'Mahyuddin Khairuddin Matyuso Nasution', or $\{w_1,w_2,w_3,w_4\}$ = {"Mahyuddin","Khairuddin","Matyuso","Nasution"} as a set of words. We obtain the power of sets $\{\{\},\{w_1\},\{w_2\},\{w_3\},\{w_4\},\{w_1,w_2\},\{w_1,w_3\},\{w_1,w_4\},\{w_2,w_3\},\{w_2,w_4\},\{w_3,w_4\},\{w_1,w_2,w_3\},\{w_1,w_2,w_4\},\{w_2,w_3,w_4\},\{w_1,w_2,w_3,w_4\}\} = \{t^{2\wedge\{k\}-1}\}$, note: 2^{k} is called k-th power of 2. In this case we have $|\{t^{2\wedge\{k\}-1}\}| = 15$ and $l = k$. Therefore, probability of t_l or $p(t_l) = 1/(2^l-1)$. Suppose the vector space of t_l is $\{t^{2\wedge\{l\}-1}\}$, 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 Ω is a set of documents indexed by search engine, i.e., a set consists of the ordered pair of the terms t_{li} and documents d_{lj}, or (t_{li},d_{li}), i=1,…,I, j=1,…,J. The relation table is two columns t_l and d_l as a representation of search engine whereby $\Omega_l = \{(t_l,d_l)_{ij}\}$ is a subset of Ω. The size of Ω is denoted by $|\Omega|$.

Definition 3. Let t_l is a search term and q is a query, then t_l in q for t_l in d_l, d_l in Ω.

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_l is true for all d in Ω: $(d => t_l) = 1$. Thus, the degree of d=>q measured by $P(d=>q)$. Therefore there are the uniform mass probability function for Ω, i.e.

$$P : \Omega \rightarrow [0,1] \tag{1}$$

where $\Sigma_\Omega P(d) = 1$.

Definition 4. Suppose t_x is a search term or t_x in S whereby S is a set of singleton search terms of search engine. A vector space Ω_x, be a subset of Ω, is a singleton search engine event (singleton) of documents that contain an occurrence of t_x in d_x.

The same meaning of Ω_x as subset of Ω is if d=>t_x has true value, or $\Omega_x(t_x)\approx 1$ if t_x is true at d in Ω or $\Omega_x(t_x) \approx 0$ otherwise, and the cardinality of Ω_x be $|\Omega_x| = \Sigma_\Omega(\Omega_x(t_x)\approx 1)$. 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_x on d=>q means that

$$P(\Omega_x) = P(\Omega_x(t_x)\approx 1) = \Sigma_\Omega(\Omega_x(t_x)\approx 1)/|\Omega| = |\Omega_x|/|\Omega|. \tag{2}$$

However, if search term in pattern, like t_x = "Mahyuddin Khairuddin Matyuso Nasution", then a different result appears. In other words, Ω_{xp}("t_x")=1 if t_x is true at d in Ω exactly or Ω_{xp}("t_x")= 0 otherwise, and the cardinality of Ω_x be $|\Omega_{xp}| = \Sigma_\Omega(\Omega_{xp}(\text{"}t_x\text{"})=1)$. In this case, each document that is indexed by search engine contains at least one occurrence of a search term. In degree of uncertainty of d=>"t_x" on d=>q is

$$P(\Omega_{xp}) = P(\Omega_{xp}(\text{"}t_x\text{"})=1) = \Sigma_\Omega(\Omega_{xp}(\text{"}t_x\text{"})=1)/|\Omega| = |\Omega_{xp}|/|\Omega|. \tag{2}$$

Thus $|\Omega_{xp}|/|\Omega| \leq |\Omega_x|/|\Omega|$, so $|\Omega_{xp}| \leq |\Omega_x|$ or Ω_{xp} is a subset of Ω_x.

Let t_x and t_y are two different search terms. If $t_x = t_y$, $t_x \neq t_y$, or $|t_x|<|t_y|$, then Ω_{xp} be a subset of Ω_x or Ω_{yp} be a subset of Ω_y or Ω_{xp} be a subset of Ω_y or Ω_{yp} be a subset of Ω_y.

3. Adaptive Approach to Model

Let t_x and t_y 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 the results of the search engine. Some of the adaptive characteristics are as follows [12, 13, 14].

Lemma 1. If $t_x \neq t_y$ and $t_x \cap t_y = \phi$, then $|\Omega_x \cap \Omega_y| = 0$ and $|\Omega_x \text{U} \Omega_y| = |\Omega_x| + |\Omega_y|$ where Ω_x and Ω_y are subsets of Ω.
Proof. $t_x \neq t_y$ and $t_x \cap t_y = \phi$ mean that for all w_x in t_x all w_x not in t_y and for all w_y in t_y all w_y not in t_x, then for all w_x in d_x all w_x not in d_y and for all w_y in d_y all w_y not in d_x such that $t_x \text{U} t_y = t_y \text{U} t_x$ and $d_x \text{U} d_y = d_y \text{U} d_x$.

Therefore, Ω_x={(t_x,d_x)} and Ω_y={(t_y,d_y)} are two independent events from queries, or t_x and t_y are true at d in Ω, respectively. In this case, $\Omega_x \cap \Omega_y = \phi$. In other words, $\{(t_x,d_x)\}U\{(t_y,d_y)\} = \Omega_x U \Omega_y$. Therefore, we have

$$|\Omega_x \cap \Omega_y| = 0 \tag{3}$$

and

$$|\Omega_x U \Omega_y| = |\Omega_x| + |\Omega_y| \tag{4}$$

Lemma 2. If $t_x \neq t_y$ $t_x \cap t_y \neq \phi$ and $|t_y|<|t_x|$, then $|\Omega_x|=|\Omega_x|+|\Omega_y|$ where Ω_x and Ω_y are subsets of Ω.
Proof. Based on assumption, we have for all w_y in t_y all w_y in t_x, but there are w_x in t_x whereby w_x not in t_y such that $t_x \cap t_y = t_y$ and $t_x U t_y = t_x$. Similar concept, for all w_y in t_y all w_y in d_y and because all w_y in t_x we conclude that w_y also in d_x, but there are w_x in t_x and x_x in d_x whereby w_x not in t_y such that w_x not in d_y. Thus, if $t_x \cap t_y = t_y$ then $d_x \cap d_y = d_y$ and if $t_x U t_y = t_x$ then $d_x U d_y = d_x$. Therefore, $\Omega_x = \{(t_x,d_x)\} = \{(t_x U t_y, d_x U d_y)\} = \{(t_x,d_x)U(t_y,d_y)\} = \{(t_x,d_x)\}U\{(t_y,d_y)\} = \Omega_x U \Omega_y$. In other words,

$$|\Omega_x| = |\Omega_x| + |\Omega_y| \tag{5}$$

Proposition 1. For search terms $t_z \neq \ldots \neq t_y \neq t_x$ and $|t_z| < \ldots < |t_y| < |t_x|$, then $|\Omega_x| = |\Omega_x| + |\Omega_y| + \ldots + |\Omega_z|$, where $\Omega_z, \ldots, \Omega_y, \Omega_x$ are subsets of Ω.
Proof. Based on generalization of Equation (3) and Equation (4), we derive $|\Omega_x| = |\Omega_x U \Omega_y| = |\Omega_x|+|\Omega_y| = |\Omega_x|+|\Omega_y U \ldots| = |\Omega_x|+|\Omega_y|+\ldots = |\Omega_x|+|\Omega_y|+|\ldots U \Omega_z|$, and

$$|\Omega_x| = |\Omega_x| + |\Omega_y| + \ldots + |\Omega_z| \tag{6}$$

Lemma 3. If $t_x \neq t_y$ $t_x \cap t_y = \phi$ and $d_x \cap d_y \neq \phi$, then $|\Omega_x| \approx |\Omega_y|$, Ω_x and Ω_y are subsets of Ω.
Proof. $t_x \neq t_y$ $t_x \cap t_y = \phi$ and $d_x \cap d_y \neq \phi$ mean that for all w_x in t_x all w_x not in t_y and for all w_y in t_y all w_y not in t_x then $t_x U t_y = t_y U t_x$, but for all w_x in d_x there are w_x in d_y and for all w_y in d_y there are w_y in d_x also, then $d_x \cap d_y = d_x = d_y$ and $d_x U d_y = d_y U d_x = d_x = d_y$. Or for $\Omega_x = \{(t_x,d_x)\}$ and $\Omega_y = \{(t_y,d_y)\}$ we obtain $\Omega_x \cap \Omega_y = \{(t_x,d_x)\} \cap \{(t_y,d_y)\} = \{(t_x,d_y)\} \cap \{(t_y,d_y)\} = \{(t_y,d_y)\} \cap \{(t_y,d_y)\}$ or $\Omega_x \cap \Omega_y = \Omega_y \cap \Omega_y$ or

$$\Omega_x \cap \Omega_y = \Omega_y. \tag{7}$$

Similarly,

$$\Omega_x \cap \Omega_y = \Omega_x. \tag{8}$$

In other words, $\Omega_x = \{(t_x,d_x)\} = \{(t_x,d_x U d_y)\} = \{(t_x,d_x)U(t_x,d_y)\} = \{(t_x,d_x)U(t_y,d_y)\} = \{(t_y,d_x)U(t_y,d_y)\} = \{(t_y,d_x U d_y)\} = \{(t_y,d_y)\} = \Omega_y$. Therefore, based on Equation (7) and Equation (8), we have $|\Omega_x| \approx |\Omega_y|$.

Definition 5. Suppose t_x and t_y are two different search terms. Let $t_x \neq t_y$, t_x and t_y in S, where S is a set of singleton search term of search engine. A doubleton search term is $D = \{\{t_x,t_y\}:t_x,t_y \text{ in } S\}$ whereby the vector space of doubleton search term denoted by $\Omega_x \cap \Omega_y$ is a doubleton search engine event of documents that contain a co-occurrence of t_x and t_y such that t_x,t_y in d_x and t_x,t_y in d_y whereby Ω_x, Ω_y, $\Omega_x \cap \Omega_y$ are subsets of Ω.

Theorem 1. Suppose t_x and t_y are two different search terms. $\Omega_x \cap \Omega_y$ is a doubleton search engine event for t_x and t_y whereby Ω_x and Ω_y are subsets of Ω, then $|\Omega_x \cap \Omega_y| \leq |\Omega_x| \leq |\Omega|$ and $|\Omega_x \cap \Omega_y| \leq |\Omega_y| \leq |\Omega|$.
Proof. Based on set theory $\Omega_x \cap \Omega_y$ be subset of Ω_x and $\Omega_x \cap \Omega_y$ be subset of Ω_y, thus $|\Omega_x \cap \Omega_y| < |\Omega_x|$ and $|\Omega_x \cap \Omega_y| < |\Omega_y|$. While based on Lemma 3 we have $|\Omega_x \cap \Omega_y| = |\Omega_x|$ and $|\Omega_x \cap \Omega_y| = |\Omega_y|$.
For $\Omega_x = \{(t_x,d_x)\}$, we have $\{(t_x \cap t_y, d_x \cap d_y)\} = \{(t_x,d_x) \cap (t_y,d_y)\} = \{(t_x,d_x)\} \cap \{(t_y,d_y)\} = \Omega_x \cap \Omega_y$ if $t_x \neq t_y$ and $|t_x|<|t_y|$. Thus $|\Omega_x \cap \Omega_y| = |\Omega_x|$. Similarly, we obtain $\Omega_y = \Omega_x \cap \Omega_y$, so $|\Omega_x \cap \Omega_y| = |\Omega_y|$. Based on Definition 5 $D = \{\{t_x,t_y\},t_x,t_y \text{ in } S\}$, or $\{t_x,t_y\} = \{(t_x,d_x),(t_y,d_y)\} = \{(t_x,d_x) \cap (t_y,d_y)\} = \{(t_x,d_x) \cap (t_y,d_y)\} = \{(t_x \cap t_y), (d_x \cap d_y)\}$

= $\{(t_x,t_y),(d_x,d_y)\}$, for $\{t_x,t_y\} = \Omega_x \cap \Omega_y$, we get $\Omega_x \cap \Omega_y = \Omega_x$ and $\Omega_x \cap \Omega_y = \Omega_y$. In other words, $\{t_x,t_y\} = \{(t_x,d_x),(t_y,d_x)\} = \{(t_x,d_x)\},\{(t_y,d_y)\}$, for $\{t_x,t_y\} = \Omega_x \cap \Omega_y$ we get $\Omega_x \cap \Omega_y = \Omega_x \cap \Omega_y$, $\Omega_x \cap \Omega_y$ is a subset of Ω_x or Ω_y. If the comma logically means "and" in set theory it means an intersection.
Therefore, $|\Omega_x \cap \Omega_y| \leq |\Omega_x| \leq |\Omega|$ and $|\Omega_x \cap \Omega_y| \leq |\Omega_y| \leq |\Omega|$ for all search terms t_x and t_y.

Corollary 1. If t_x and t_y are the different search terms, then $|\Omega_x \cap \Omega_y| = |\Omega_x \cap \Omega_y| + |\Omega_x \cap \Omega_x| + |\Omega_y \cap \Omega_y|$.
Proof. As the direct or indirect consequence of Proposition 1 and Theorem 1.

4. A Selective Approach as Simulation

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 Afif Siregar, Achsanuddin Hanafie, Adi Kusuma Aman, Alfred C. Satyo, Askaroellah Aboet, Atan Baas Sinuhaji, Ayodhia Pitaloka Pasaribu, Aznan Lelo, Bachtiar Surya, Budi R. Hadibroto, Chairuddin Panusunan Lubis, Chairul Yoel, Darwin Dalimunthe, Daulat Hasiholan Sibuea, Delfi Lutan, Delfitri Munir, Erwin Dharma Kadar`}. Among the names of actors as the term, two different terms t_x and t_y have several options that 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.

Table 1. Experiment design for simulation of search engine

Actor Name (A)	Medium of randomness test		Search engine as test simulation			Search engine as comparative simulation		
	1	2	$\|\Omega_x\|$	$\|\Omega_{"x"}\|$	$\|\Omega_x \cap \Omega_y\|$	$\|\Omega_x\|$	$\|\Omega_{"x"}\|$	$\|\Omega_x \cap \Omega_y\|$
a	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
…	…	…	…	…	…	…	…	…
b	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
z	…	….	…	…	…	…	…	…
Average	pr or lc	a or n	avg_1	avg_2	avg_3	avg_4	Avg_5	avg_6
n_1	n_1(pr)	n_1(a)	$n_{11}(\geq)$	$n_{12}(\geq)$	$n_{13}(\geq)$	$n_{14}(\geq)$	$n_{15}(\geq)$	$n_{16}(\geq)$
n_2	n_2(lc)	n_2(n)	$n_{21}(<)$	$n_{22}(<)$	$n_{23}(<)$	$n_{24}(<)$	$n_{25}(<)$	$n_{26}(<)$
Run (r)	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
μ_r	μ_{r1}	μ_{r2}	μ_{r3}	μ_{r4}	μ_{r5}	μ_{r6}	μ_{r7}	μ_{r8}
σ_r	σ_{r1}	σ_{r2}	σ_{r3}	σ_{r4}	σ_{r5}	σ_{r6}	σ_{r7}	σ_{r8}

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 columns contain the list of singletons respective to t_x and t_x in quotes, and a list of doubletons of t_x and t_y (singleton with keyword). In this case, we ensure that the singletons also are random.

IOP Conf. Series: Journal of Physics: Conf. Series **801** (2017) 012078 doi:10.1088/1742-6596/801/1/012078

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 seperate the sample into two categories: number of first categories

$$n_1 = \sum_{i=1\dots n} a_{i1} \tag{1}$$

or

$$n_2 = \sum_{i=1\dots n} a_{i2} \tag{2}$$

whereby a_{i1} is elements of A that meet first category and a_{i2} is elements of A that meet second category. While run (r) is how many times the category change in the sample. Thus, the average of run is

$$\mu_r = (2n_1n_2/(n_1+n_2))+1 \tag{3}$$

and the variance of run is

$$\sigma_r^2 = ((2n_1n_2(2n_1n_2-n_1-n_2))/((n_1+n_2)^2(n_1+n_2-1)))^{1/2}. \tag{4}$$

Then, we have Z_{count} as follows

$$Z_{count} = (r - \mu_r)/\sigma_r \tag{5}$$

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 $n_1 = 34$ and $n_2 = 17$. By using Equations (3), (4), and (5), we obtain $\mu_r = 23.67$, $\sigma_r = 0.93$ and $Z_{count} = -1.79$, and for $\alpha = 0.05$ we obtain $Z_{\alpha=-0.025}=1.96 \leq Z_{count} \leq Z_{=0.025} = 1.96$, and because r is located between the critical value then the decision is received H_0. 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 (χ^2). Suppose the data space (ds) is presented in matrix form as follows,

$$ds = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{vmatrix}$$

Amount of data x_{ij} is S_{ij} as follows

$$S_{ij} = \Sigma_{i=1,\dots,n,j=1,\dots,m}\, x_{ij}\ . \tag{6}$$

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

$$\begin{aligned} e_{11} &= (\Sigma_{i=1,j=1,\dots,m}\, x_{ij})(\Sigma_{i=1,\dots,n,j=1}\, x_{ij})/S_{ij} \\ e_{12} &= (\Sigma_{i=1,j=1,\dots,m}\, x_{ij})(\Sigma_{i=1,\dots,n,j=2}\, x_{ij})/S_{ij} \\ &\dots \\ e_{mn} &= (\Sigma_{i=n,j=1,\dots,m}\, x_{ij})(\Sigma_{i=1,\dots,n,j=m}\, x_{ij})/S_{ij} \end{aligned} \tag{7}$$

and we have a matrix of expectations as follows

$$es = \begin{vmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ e_{21} & e_{22} & \dots & e_{2n} \\ \dots & \dots & \dots & \dots \\ e_{m1} & e_{m2} & \dots & e_{mn} \end{vmatrix}$$

Amount of data e_{ij} is E_{ij} as follows

$$E_{ij} = \Sigma_{i=1,\ldots,n,j=1,\ldots,m}\ e_{ij} \tag{8}$$

Then, we have

$$\chi^2 = \Sigma_{i=1,\ldots,n,j=1,\ldots,m}\ (x_{ij}\text{-}e_{ij})/e_{ij} \tag{9}$$

with degree of freedom (df) is (m-1)(n-1). For example, among 51 actor names we have x_{11} = 34 professors, x_{21} = 17 lectures, x_{12} = 17 authors, and x_{22} = 34 non-authors. Based on Eq. (7) we can calculate their expectations, i.e. $e_{11} = e_{12} = e_{21} = e_{22} = 25.5$, and based on Eq. (9) we obtain $\chi^2 = 11.33$ 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 3.841, then rejects the null hypothesis of independence because $\chi^2 > 3.841$. This tell us there is a relationship between type of academic level and authors.

In 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_x, $t_{\text{"}x\text{"}}$, and t_x,t_y by completing the computations as follows:
    a. For t_x, we have the amount of 51 $|\Omega_x|$ from Google search engine, that is 2635514 with average (avg) = 51676.75. Number of $|\Omega_x|$ greater than or equal to avg is 10 $|\Omega_x|$ while number of $|\Omega_x|$ less than avg is 41. By using Eq. (3), Eq. (4), and Eq. (5), Z_{count}=-6.54 < $Z_{\alpha=-0.025}$=-1.96. Therefore, reject H_0 and 51 singletons of t_x from Google search engine is not random.
    b. However, the amount of 51 $|\Omega_{\text{"}x\text{"}}|$ from Google search engine, that is 1095045 with average (avg) = 21471.47. Number of $|\Omega_{\text{"}x\text{"}}|$ greater than or equal to avg is 3 $|\Omega_{\text{"}x\text{"}}|$ while number of $|\Omega_{\text{"}x\text{"}}|$ less than avg is 48. By using the similar equations, Z_{count}=-1.50 > $Z_{\alpha=-0.025}$=-1.96, and H_0 accepted whereby 51 singletons of $t_{\text{"}x\text{"}}$ from Google search engine is random.
    c. For doubleton t_x,t_y whereby t_y = "Universitas Sumatera Utara" as a keyword, we have amount of 51 $|\Omega_x \cap \Omega_y|$ from Google search engine, i.e 61092 with avg = 1197.88. Number of $|\Omega_x \cap \Omega_y|$ greater than or equal to avg is 15 and number of $|\Omega_x \cap \Omega_y|$ less than avg is 36. With that, we obtain Z_{count}=-1.31 > $Z_{\alpha=-0.025}$=-1.96 based on Eq. (3), Eq. (4) and Eq. (5), and H_0 accepted, thus 51 doubletons of t_x,t_y from Google search engine is random.
    d. Whereas, for t_x by using Yahoo search engine, we have the amount of 51 $|\Omega_x|$ is 2365061 with avg is 46373.76. So $n_1(pr)$ = 5 and n_2(lc) = 46. Z_{count}=-0.03 > $Z_{\alpha=-0.025}$=-1.96. On that basis, H_0 accepted, thus 51 singletons of t_x from Yahoo search engine is random.
    e. Simlarly for $t_{\text{"}x\text{"}}$, the amount of 51 $|\Omega_{\text{"}x\text{"}}|$ from Yahoo search engine, that is 395815 with average (avg) = 7761.08. Number of $|\Omega_{\text{"}x\text{"}}|$ greater than or equal to avg is 3 $|\Omega_{\text{"}x\text{"}}|$ while number of $|\Omega_{\text{"}x\text{"}}|$ less than avg is 48. By using the similar equations, Z_{count}=-1.50 > $Z_{\alpha=-0.025}$=-1.96, and H_0 accepted whereby 51 singletons of $t_{\text{"}x\text{"}}$ from Yahoo search engine is random.
    f. For doubleton t_x,t_y whereby t_y = "Universitas Sumatera Utara" as a keyword, we have amount of 51 $|\Omega_x \cap \Omega_y|$ from Yahoo search engine, i.e 15361 with avg = 301.19. Number of $|\Omega_x \cap \Omega_y|$ greater than or equal to avg is 12 and number of $|\Omega_x \cap \Omega_y|$ less than avg is 39. With that, we obtain Z_{count}=-0.42 > $Z_{\alpha=-0.025}$=-1.96 based on Eq. (3), Eq. (4) and Eq. (5), and H_0 accepted, thus 51 doubletons of t_x,t_y 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

Categories	Samples					
	Google search engine			Yahoo search engine		
	$\|\Omega_x\|$	$\|\Omega_{"x"}\|$	$\|\Omega_x \cap \Omega_y\|$	$\|\Omega_x\|$	$\|\Omega_{"x"}\|$	$\|\Omega_x \cap \Omega_y\|$
n_1	10	3	15	5	3	12
n_2	41	48	36	46	48	39

   a. First, we test the independence $|\Omega_x|$ of Google and $|\Omega_x|$ of Yahoo. By using Eq. (6), Eq. (7), Eq. (8), and Eq. (9) toward $n_1(|\Omega_x|)$ and $n_2(|\Omega_x|)$ 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 $|\Omega_{"x"}|$ of Google and $|\Omega_{"x"}|$ of Yahoo. By using similar equations against $n_1(|\Omega_{"x"}|)$ and $n_2(|\Omega_{"x"}|)$ 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 $|\Omega_x \cap \Omega_y|$ of Google and $|\Omega_x \cap \Omega_y|$ of Yahoo. By using similar equations with $n_1(|\Omega_x \cap \Omega_y|)$ and $n_2(|\Omega_x \cap \Omega_y|)$ see Table, we get value of $\chi^2 = 0.45 < 3.84$ for $df = 1$, and H_0 accepted for $\alpha = 0.05$. Therefore, two samples are independent.
   d. For getting behavior of $|\Omega_x|$, $|\Omega_{"x"}|$, and $|\Omega_x \cap \Omega_y|$, we test independence among singletons and doubleton of Google search engine. By using Eq. (6), Eq. (7), Eq. (8), and Eq. (9) for $n_1(|\Omega_x|)$, $n_1(|\Omega_{"x"}|)$, $n_1(|\Omega_x \cap \Omega_y|)$, $n_2(|\Omega_x|)$, $n_2(|\Omega_{"x"}|)$, and $n_2(|\Omega_x \cap \Omega_y|)$ see Table 2, we obtain $\chi^2 = 9.53 > 7.82$ with $df = 3$, 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 doubletons 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 Ω has a normal distribution, where Eq. (1) 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 doubleton 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.

IOP Conf. Series: Journal of Physics: Conf. Series **801** (2017) 012078 doi:10.1088/1742-6596/801/1/012078

5. Conclusion

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: Adaptive properties for doubleton *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*.