A Proposed Method to Recognize the Research Trends using Web-based Search Engines

Mohammed R. Elkobaisi
Department of Computer
Omer AL-Mukhtar University, Libya
elkobaisi@gmail.com

Abdelsalam M. Maatuk
Faculty of Information Technology
Benghazi University, Libya
abdelsalam.maatuk@uob.edu.ly

Shadi Aljawarneh
Software Engineering Department
Jordan University of Science and Technology, Irbid, Jordan

ABSTRACT
This paper presents a novel approach to recognize research trends in a particular domain of research (i.e. Agent development) that is based on the number of data extracted from search engines. Several well-known mathematical and statistical theories have been used, from which a mathematical model has been derived to predict the agent development. The proposed solution attempts to convert the raw resulted number of documents found on the Internet (in PDF format) into useful information, and fit a curve representing the number of searched data of every year. A prototype has been designed, to search multi-keywords at the same time and collect the required data automatically to eliminate the useless data, before converting it into usable data.

Keywords
Search Engines, Information Extraction, Agent systems, Software Engineering.

1. INTRODUCTION
The information available online today is growing rapidly, so that the main challenge is how to extract real volume of data in a specific field. The resources of data are spreading and easy to obtain them. Search engines are designed to search for information on the Internet and FTP servers, and use regularly updated indexes to operate efficiently [10]. A search engine is a program that searches documents for specified keywords and returns a list of the documents where the keywords were found. The search engine allows one to ask for a content meeting specific criteria and retrieving a list of files that match those criteria [11]. When a query submitted into a search engine, it examines its index and provides a set of best-matching pages according to its criteria, with information about the documents that user needs. The most popular engines currently available are Google and Yahoo [10].

Information extraction from search engine results is a promising approach to grasp the research trends. Search engines mine data available in large databases or open directories [10]. This creates opportunities for new techniques to collect the required data from the Internet.

ICEMIS '15, September 24-26, 2015, Istanbul, Turkey © 2015 ACM. ISBN 978-1-4503-3418-1/15/09$15.00
DOI: http://dx.doi.org/10.1145/2832987.2833012

There are many ways used in search engines to search for documents based on keywords relevant to a specific topic. Moreover, prototype systems can be developed as tools incorporated with common research engines in order to use the collected data from such engines. These tools can inspect and analyse data for research publications in any field of knowledge to recognize the challenges facing the developing research areas.

This paper presents a method which recognizes research trends based on the number of data by search engines. It argues the challenges that face the development of agent and agent research information. An agent is a computational system that sense and act autonomously in a complex dynamic environment to satisfy a set of goals or tasks for which it is designed [12]. The collected raw data is processed using mathematical and statistical theories to convert it into a useable data. A mathematical model is formulated based on several well-known mathematical and statistical theories to predict the agent development. The method attempts to fit a curve to the number of searched data of every year. The information format is extremely varied; however, we focus on documents in PDF format, as it becomes the most popular standard for document publishing. The actual noisy data are analysed and computed from the collected data to reduce the noise affects. A prediction system is designed, based on mathematical models for researching topics and find out the number of their PDF documents stored in the Internet.

This paper is structured as follows. Section 2 reviews briefly the related work. An overview of the proposed method is presented in Section 3. Section 4 concludes the paper.

2. RELATED WORK

With the rapid development of internet and the continuous increase of the information, introducing the web extraction technology into traditional search engines becomes one of the most important challenges in information retrieval and artificial intelligence fields. Many studies have been proposed on web search result extraction and the trend of web-based search engines [10, 11, 13, 17]. Existing approaches in automatic web extraction can be classified into three categories, which are: wrapper, automatic template generation and a hybrid approach [17]. Earlier works in information extraction are mainly semi-automatic or even manual [14, 15], which are relied on training and human assistance to generate extraction rules for web pages. Many new applications such as building large-scale meta-search engines [16] require fully automated wrapper generation techniques. However, existing work does not appear to provide a solution for using the resulted number of search query to predict the development trends of the research in a specific field.

Compared to existing approaches, we emphasis on information retrieved from search engine not on the search engine itself. We propose a new method, which detects automatically results for searching of multi-keywords without user interaction and manual process.

3. OVERVIEW OF PROPOSED METHOD

The method described here uses the Internet's search engines as a tool for data collections. The interested results of the search process are the number of items found for pre-defined requirement searched using the search engine. This requirement is defined as a set of keywords and logical relations (e.g., "or", "and") among them and constraints on them. Note that the data source of this study is the Internet. The collected data require empirical studies to apply the idea and predict the empirical formula for the phenomena under study. Each and every search engine has a specific set of rules that can be used to define the user requirement. This method of data collection is different from the others, such as questionnaire and interviews. These methods require different treatment to extract the needed data for study. Data is collected in this research as follows:

1. The keywords and their relation structure are defined firstly.
2. The defined keywords and relation structure is fed to the selected search engine, then the search engine is executed to search for the required task, which the number of items found.
3. The result number is recorded against selected independent variables.
4. The recorded data is treated by using proper number of methods to extract information and knowledge needed.

The resulted data stored in a text file that contains the number of searched items, which satisfy the required keywords and their relation years as independent variables. Based on the internet usage in research work publication, it can be used as a tool to collect and monitor data about the research publications in any field of knowledge. This study is based on the following assumptions:

1. The number of research papers that are in PDF file format is proportional to the real number of papers and publications in some certain fields. Increasing the publications in a field increases the number of PDF files in the same field.
2. The percentage of the number of PDF papers to the number of PDF format papers in the interval of years is proportional to the percentage of the real publications in that specified year to the specified interval of years for the same subject.
3. Growth function and other mathematical functions are used to model the change of publications number, where some characteristics can be retrieved using such functions.

3.1 Problem Definition

Because an agent system [5] is a new and evolving rapidly, this might arise some problems related to this important discipline of software engineering, which can be formulated as questions:

- What are the best practice methods to measure the development of agent?
- What are the trends of academics agent-researches (both successful and unsuccessful) in past, current and future?
- What are the major challenges that affect agent deployment?
- How can we imagine the reasons for success and failure of some agent-aspects in a period of time?
- How can we help industries to protect their investments and smoothly evolve agent-based services, solutions, systems and products?

The majority of Internet search users use Google and Yahoo engines. Google still appears to have the largest database of Web pages, including many other types of Web documents. Yahoo is getting the second rank of search engines [1]. The data for this work is collected through the internet via search engine using suitable keyword based on a hypothetical idea about the relation between the number of published papers and their corresponding number of PDF files found. The keywords and their relational structure are designed to represent the required information. These keywords have the general structure: "**Keyword**" + "**Year**" + **PDF**, where: **Keyword** is the searching martial; it may be one word or many words. **Year** is the year of publications, and **PDF** is the type of data format [6]. Figure 1 shows the headers of Google and Yahoo engines after search process, the used keywords in the textbox and the resulted number of items in marked places.

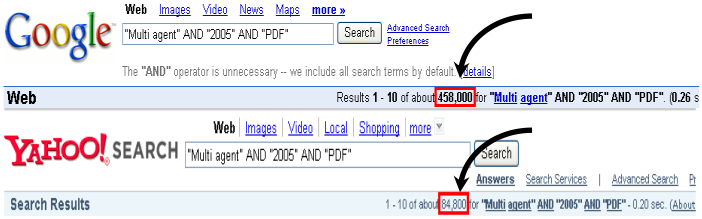

Figure 1. The resulted number of search engines

3.2 Mathematical background and basic model

We have derived our model from a number of mathematical formulas. The mathematical model is used to represent the collected data by fitting it to the mathematical formula. A mathematical model is found by computing the best coefficients, and then it can be used to trace the behaviour of the system, which gives the collected data. The mathematical expressions are used as model that represents the collected data and given in the following subsections.

3.2.1 Growth function and activity growth

Research activity in a certain area is starting with single or some of the basic research that guide the research to unsolved problems. The old points open the gate to other new points, so the number of research work in that area will grow exponentially. This type of growing is represented mathematically using a formula that is known as a mathematical growth function or an expression. In this type of mathematical expression, the increase of the number of items is depending on the actual number before the growth process. That is:

$$N = N_0(1+a)^x \qquad (1)$$

Where N_0 is initial number, N is new number after an interval, x is the length of interval and a is the growth rate per interval. The differential of such system is:

$$\frac{dN}{dt} = \alpha N \qquad (2)$$

Note that this expression for growth rate or the change of the growth with time and depending on α growth factor and the actual number of items. If *N* represents the number of research work at a certain time, then the new number of researches can be found if the growth rate is known, by using Euler Method [7] for solving numerical differential equation of the form:

$$N = N_0(1 + h\alpha) \qquad (3)$$

This is of one interval, if the new number is used as initial value for the coming interval, then the new forms as:

$$N = N_0(1 + h\alpha) \ /(1 + h\alpha) \qquad (4)$$

3.2.2. Steady state activity

Any balanced system reaches the steady state due to the interaction with other factors that work against its growth in unlimited case. The balance between interacting elements leads to this steady state or balance, in which the rate of growth is nearly equal to the rate of decreasing due to the other factors. This type of system can be expressed in the mathematical expressions given below:

$$\frac{dN}{dt} = A_1 N - B_1 M \qquad (5)$$

$$\frac{dM}{dt} = A_2 N - B_2 M \qquad (6)$$

In such system, the growth of *N* is depending on the change of *N* and the change of *M* values. The two variables are working against each other. In the case of our study, the research work is increasing with the increase of the number of research work, but this decreases by the effect of the obstacles found in research work. If the increasing and decreasing rates are equal, then the rate of growth is equal or nearly equal to zero, and it will stay on fixed level until something changes the balance. Because of this, it decreases or increases depending on the direction of growth rate, either positive or negative.

3.2.3 Moving between two steady states

The balanced system at steady state stays unchanged in its current state until a new event occurs, that will force it to change its state to new one either higher or lower. In general, the change of the state has "ς" shape or its reverse shape of "ς", so that it can be modelled using Sigmoid function formula [8] as follows:

$$y = \frac{1}{1 + e^{-ax}} \qquad (7)$$

If *y* is dependent variable or output, *x* is independent variable or input, as *x* is changed from (-∞ to ∞), and the value of *y* between (0 and 1). A range of *y* can be changed to any other range by using the following formula:

$$y_m = (y_1 - y_2)\left(\frac{1}{1 + e^{-aX}}\right) + y_0 \qquad (8)$$

Where y_0 is the minimum value and y_1 is the maximum value. In the research process, there are many stages, in which the activity is moving from one steady state to another. The first steady stage is the basic research stage that represents the interest in new field research work. After the main stage, next four stages are: basic, implementation, industry, and Laggards [2]. There are some other research stages in between appear due to some problems that work against the growth of research in the field, these stages can be considered as deflection points in research work history. Such points are responsible for redirecting research activity to overcome some problems appear in research growth path. Figure 2 shows a multi-stages system that can be represented by a mathematical expression composed of many Sigmoid functions [9], which can be expressed as:

$$y_1 = y_{m1} + y_{m2} + ... + y_{mn} = \sum_{n=1}^{N} y_{mn} \qquad (9)$$

$$= \sum_{n=1}^{\infty}\left((y_{1n} - y_{0n})\left(\frac{1}{1 + e^{-anX}}\right) + y_{0n}\right) \qquad (10)$$

The change of state can be expressed as a peak function, which is Sigmoid function differentiation.

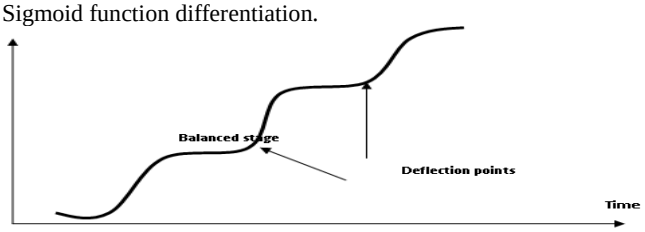

Figure 2. Multi stages system

3.3 Numerical and statistical methods

A number of numerical and statistical methods used, which are needed to predicate the relations and find the parameters of the best fitting of the resulted data to the designed models. These methods are described as follows:

3.3.1 Correlation coefficient

Correlation coefficient [3] is a computed factor represents the strength of the relation between two variables; its value larges between -1 and 1. The stronger relation has absolute value near to one, while the weak relation value is near to zero. The correlation coefficient $\rho X, Y$ between two random variables *X* and *Y* with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho X, Y = \frac{\mathrm{cov}(X, Y)}{\sigma X \sigma Y} = \frac{E((X - \mu X)(Y - \mu Y))}{\sigma X \sigma Y} \qquad (11)$$

Where *E* is the expected value operator and **cov** means covariance. Since $\mu X = E(X)$, $\sigma X2 = E(X2) - E2(X)$ and likewise for *Y*. The correlation is defined only if both of the standard deviations are finite and both of them are non-zero. A correlation between two variables is the measurement of error around estimates of one or both variables, in which case attenuation provides a more accurate coefficient [3].

3.3.2. Least Square Method

Least square method [9] is a statistical approach to estimate an expected value or function with the highest probability from the observations with random errors. The highest probability is replaced by minimizing the sum of square of residuals in the least square method. Residual is defined as the difference between the real value and an estimated value of a function. Least square method is commonly applied for the case measurements (X_i , Y_i)

are given, the relationship between x and y is estimated by a function, for example liner relation: $y = f(X) = aX + b$. By minimizing the square sum of residuals, the unknown parameters a and b will be determined [4]. Unknown a and b parameters in the case of $aX + b$ are determined as follows:

1. An observed set of equations; in matrix form is
$$XA=Y \qquad (12)$$
Where X_i and Y_i are matrices of the observed points, A is the matrix of coefficients.
2. Multiply both sides of the resulted matrix equation by transpose matrix of X (X^t), we get:
$$X^t XA = X^t Y \qquad (13)$$

The resulted matrix XX^t is a square matrix, its inverse can be computed and multiply by both sides of the above equation, so that the result will be :
$$A = (X^t X)^{-1} X^t Y \qquad (14)$$
This is an unknown coefficient matrix. The computed value $(x^t x)^{-1} x^t$ is known as a generalized computed result in matrix form for the known matrix for line case is:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum X^2 & \sum X \\ \sum X & N \end{bmatrix} \begin{bmatrix} \sum XY \\ \sum Y \end{bmatrix} \qquad (15)$$

N is the number of data points. This idea can be generalized to any higher order polynomial or any multi-independent variables. Non-linear relation needs to be linear by mathematical relation or approximated as linear to get the coefficients.

3.3.3 Polynomial interpolation

For a given point n there exists a unique polynomial of degree $(n-1)$, which passes through these points. A continuous function $f(X)$ of a polynomial of degree n may be used to represent the 'n+1' data values. The polynomial will pass through all given points. This polynomial can be used to compute the approximate values of any points within the given interval of data. If 'X' falls outside the range of given data is known as extrapolation. For four points interpolating polynomial of degree 3 and can be found by using these step formulae. The points are (X_0, Y_0), (X_1, Y_2), (X_3, Y_3), which can be written as in the general formula:

$$Y = \sum_{i=0}^{N} \frac{\prod_{\substack{i=1 \\ i \neq j}}^{N} (X - X_i)}{\prod_{\substack{i=1 \\ i \neq j}}^{N} (X_j - X_i)} Yi \qquad (16)$$

This will compute the proper coefficients of a polynomial of N degree that will pass through all given points of data.

4. CONCLUSION

We have proposed a solution to use fairly raw information from a range of search engines to examine trends in a particular domain of research, i.e., Agent research development. In addition, a mathematical model has been formulated to explore agent research behavior and its aspects and arguing the challenges that this filled faces. The approach intends to fit a curve to the number of searched data for every year. In a part of future work, the prototype will be implemented for research topics on PDF documents published on the internet. In addition, the proposed method could be improved to deal with different types of documents, such as doc, images, and video.

5. REFERENCES

[1] Chris A. 2014. *Top 10 search engines in the World*. [Online]. Available:https://www.reliablesoft.net/top-10-search-engines-in-the-world/.

[2] Levitt T. 1965. *Exploit the Product Life Cycle*. Harvard Business Review, vol. 43(6), pp. 81–94.

[3] Rodgers J. and Nicewander W. 1988.Thirteen ways to look at the correlation coefficient. In *The American Statistician*. vol. 42(1), pp. 59-66.

[4] Gerald W. 2001. *Numerical Methods with MATLAB: Implementations and Applications*. [Online]. Available: https://www.prenhall.com/recktenwald

[5] Ferber J. 1999. *Multi-agent systems: An introduction to distributed artificial intelligence*. Addison Wesley Longman, Harlow, UK.

[6] Google Advance search settings. [Online]. Available: https://www.google.com.ly/advanced_search?hl=en

[7] Atkinson K., Han W., and Stewart D. E. 2011. *Numerical Solution of Ordinary Differential Equations*. John Wiley & Sons Inc.

[8] Han J. and Claudio M. 1995. The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning. In *Proceedings of IWANN'96*. Springer-Verlag, UK, pp. 195-201.

[9] Chavent G. 2010. *Nonlinear Least Squares for Inverse Problems*. Theoretical Foundations and Step-by-Step Guide for Applications. Springer, Netherlands.

[10] Kuyoro O., Okolie O., Richmond U. and Awodele O. 2012. Trends in Web-Based Search Engine. *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3(6), pp. 942-948.

[11] Gupta R., Suri N., Bhatnagar H. and Ivy B. 2012. Find Tail - A Search Engine. *International Journal of Emerging Technology and Advanced Engineering*, vol. 2(6).

[12] Burgin, M. and Dodig-Crnkovic, G. 2009. A Systematic Approach to Artificial Agents, *In Computer Science, cs. AI*. [Online] Available: http://arxiv.org/abs/0902.3513

[13] Kuyoro O., Okolie O., Kanu U, and Awodele O. 2012. Trends in Web-Based Search Engine. In *Journal of Emerging Trends in Computing and Information Sciences*. vol. 3(6).

[14] Baumgartner R., Flesca S. and Gottlob G. 2011. Visual web information extraction with Lixto. In *VLDB '01 Conference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 119-128.

[15] Hsu C. and Dung M. 1998. Generating finite-state transducers for semi-structured data extraction from the Web. In *Information Systems*. vol. 23(8), pp. 521-538.

[16] Meng W., Yu C. and Ozsu M. 2010. Advanced Meta search Engine Technology. *Application in the Morgan and Claypool series*, E-book 1st edition.

[17] Gozudeli Y., Karacan H., Yildiz O., Baker M., Minnet A., Kalender M. and Akcayol M. 2015. A New Method Based on Tree Simplification and Schema Matching for Automatic Web Result Extraction and Matching. In *Proceedings of the International Multi Conference of Engineers and Computer Scientists*. vol. 1.