Structural convergence of web graph, social network and malaria network: an analytical framework for emerging web-hybrid search engine

Monday Eze*

Department of Computer Science/Maths/Statistics/Informatics,
Federal University,
Ndufu-Alike, Ikwo (FUNAI), Ebonyi State, Nigeria
E-mail: eze_monday@yahoo.com
*Corresponding author

Jane Labadin and Terrin Lim

Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak,
94300 Kota Samarahan, Malaysia
E-mail: ljane@fit.unimas.my
E-mail: terrin@fit.unimas.my

Abstract: It is a well known truth that there is no malaria transmission without blood-sucking bites by mosquitoes. Since there is no blood-sucking without contacts, it follows that the contacts between human beings and mosquitoes can be exploited to generate contact networks. Active computational research in this area is geared at developing new frontiers in modelling malaria transmission through network theory and methods. Ongoing research in this area points to the fact that there are structural similarities between web graph, social network and malaria networks. It is the aim of this paper to explore the structural convergence and to exploit this to build a framework that will pave the way for developing web-hybrid search engines. Like the emergence of search engines revolutionised web research, it is expected that application of search engines in malaria research will make a tremendous impact in malaria control.

Keywords: structural convergence; web graph; social network; malaria network; web-hybrid search engine; PageRank; HITS.

Reference to this paper should be made as follows: Eze, M., Labadin, J. and Lim, T. (2014) 'Structural convergence of web graph, social network and malaria network: an analytical framework for emerging web-hybrid search engine', *Int. J. Web Engineering and Technology*, Vol. 9, No. 1, pp.3–29.

Biographical notes: Monday Eze is currently a Lecturer/Academic Adviser in the Department of Computer Science/Maths/Statistics/Informatics, Federal University, Ndufu-Alike, Ikwo (FUNAI), Ebonyi State, Nigeria. He is a graduate of Computer Science. He holds an MBA in Management and an MSc in Computer Science. He received his PhD in Computer Science from the Universiti Malaysia, Sarawak. He has over 12 years of professional computing experience, which were spent working mostly in the banking industry. He is a full member of the Nigerian Computer Society (NCS). His research interest includes computational algorithms, heterogeneous contact network modelling, collaborative e-learning, business computing, etc.

Jane Labadin is currently an Associate Professor at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS). She received her PhD in Computational Mathematics specialised in Fluid Dynamics from the Imperial College of Science, Technology and Medicine, London, UK. Her main research contribution ranges from the field of population dynamics to the field in fluid dynamics. Her research interest is generally in mathematical and computational modelling of dynamical systems. Apart from research and graduate supervisions, she is also teaching advanced undergraduate courses for the computational science programme.

Terrin Lim is attached to the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak as a Lecturer. He has been teaching for the past eight years in various undergraduate courses in the Department of Computational Science and Mathematics. His research areas of interest are in statistical modelling in bioinformatics and 3D face recognition, network modelling in disease transmissions and he is currently undertaking research on network models in printed historical documents. He graduated with an MSc in Bioinformatics and Bachelor degree in Computer Science.

1 Introduction

Malaria is one of the most dangerous and widest spread tropical diseases (Global Risk Forum, 2009). It is a vector-borne disease that results from blood infection by protozoan parasites of the genus *plasmodium*, which are transmitted from one human being to another by female *anopheles* mosquitoes (Richard and Kamini, 2002). The four species of malaria parasites that infect humans are *Plasmodium falciparum, Plasmodium vivax, plasmodium malariae* and *plasmodium ovale*. Since malaria is a vector-borne disease, it follows that any research effort geared towards eradicating the disease could be termed incomplete if it ignores the prominent role of the vector in disease transmission. It is therefore not strange to hear such statements as "*Imagine there are no mosquitoes, then there will be no itch or scratch and no malaria*" as was commented by Jonathan (2011).

The importance of vectors in malaria transmission has led to a number of scientific researches geared towards tackling malaria eradication from the angle of vector control. One of such researches is the genetic alteration of mosquito DNA and subsequent introduction of large number of impotent males into wild populations. For instance, the Malaysian Government through its National Biosafety Board recently approved the release of genetically modified (GM) mosquitoes in some selected forests (Shiow, 2010). This is the second in history after the Cayman Islands (Ethan, 2010) which took place in 2009. About 6,000 GM mosquitoes, known as OX513A, developed by Malaysia's Institute for Medical Research (IMR) in collaboration with Oxitec (a UK-based biotech company) were released, with the hope that the male GM mosquitoes will compete with normal males for females. The expected effect is a drastic reduction in number of *A. aegypti* which is the vector that transmits dengue.

A sophisticated technology currently in its infant stage of development is the laser-wielding robot that can detect mosquitoes in the air and shoot them dead at an enormous rate of hundreds of pests per second. Two serious concerns expressed from public opinion is the fact that the laser could harm human beings, and secondly, the fear that the technology may not be smart enough to avoid hitting other insects which are useful in the ecosystem. Remote sensing and satellite technology have also been exploited by scientists to predict hotspots of malaria vectors as mentioned by Timothy et al. (2009), Charoenpanyanet and Chen (2008), Satya et al. (2007) and Naoko et al. (2002).

It is pertinent to mention that application of scientific tools and deployment of technologies to control vectors will be very ineffective if there are no precision techniques for detecting locations where these vectors are hiding. In other words, scarce resources could be wasted trying to control vectors where they do not exist, while neglecting areas of high concentration. Literature survey has shown a number of weaknesses with some existing techniques for predicting malaria vectors hotspots. The quest for new computational techniques that will improve on existing methods in this important area of malaria research is the motivation for proposing a network-based approach to malaria reservoir detection. Some of the weaknesses of existing techniques for detecting vector reservoirs will be outlined here.

First, the American Mosquito Control Association (AMCA, 2011) reported the anomaly commonly known as 'mistaken identity', a situation where several insects are usually mistaken as mosquitoes. As many as five insects were enumerated in this category: crane flies, midges, fungus gnats, dance flies, and wood gnats. One serious concern which cannot be ignored is as follows. Could the outcome of using satellite technology and remote sensing to predict the malaria vectors reservoirs be prone to error resulting from mistaken identity? While this important question remains unanswered, the probability of such errors cannot be ignored either.

Second, it has been reported that there are over 3,000 species of mosquitoes, of which about 100 are vectors of human diseases, with only about 30 (which is 1% of 3,000) so far known to transmit *Plasmodium*, the parasite that cause malaria (Abirami et al., 2011). Here are two possible challenges that are expected to confront the existing techniques. Apart from the challenge of identifying the 1% of mosquitoes species which cause malaria out of over 3,000 existing species, there is a likely challenge of being able to identify the infected mosquitoes population. This is because only the infected mosquitoes are actually responsible for malaria transmission. This task may not be very easy to achieve using existing techniques. In fact, the remote sensing images seen so far only estimate the vector densities, but fail to further classify such outputs in terms of infected vector populations. For instance, though it is stated that there are currently about 80 species of mosquitoes in Florida (FMLAB, 2011), USA (and in particular Florida) is not recognised as a malaria endemic area. Hence the existences of mosquitoes in an area do not necessarily imply the existence of malaria, however, the existence of malaria always imply the existence of infected mosquitoes in the particular location.

Based on the issues discussed so far, our ongoing research proposes to apply computational modelling with the aim of discovering not just the vector reservoirs, but the ones that habour the infected mosquitoes. The model is based on the contact network formed as a result of interactions between the human beings who have had malaria incidences and the infected mosquitoes. A common assumption in classical infectious disease modelling methods such as SIR (susceptible-infected-recovered) model is population homogeneity, where all the population is assumed to be made up of groups that mix uniformly at random. This assumption has been described as unrealistic by Tom and Gerardo (2009). Rather than assuming that every infectious individual (regardless of age, geographic location, etc.) has the same probability of coming in contact with any susceptible individual in the population, a model closer to reality incorporates the fact that the population is heterogeneous, and is characterised with varied contact rates, among other attributes (Tom and Gerardo, 2009). Network modelling is one method that takes into consideration the heterogeneous nature of the interacting components in disease transmission.

This paper is a framework for developing a search engine that discovers the reservoirs of infectious malaria vectors using contact network modelling. It is based on the fact that every malaria transmission involves malaria vector contacts with human beings through mosquito bites. The proposed search engine ranks the public places based on the density of infected malaria vector population. The public places are locations, other than residential homes, which human beings visit for their socio-economic activities (e.g., schools, markets, football pitch, etc). To successfully implement the proposed network-based search engine, there must be a network which guarantees the search space. From previous research in this area, a malaria network has been evolved, and its structural attributes well defined by Eze et al. (2011c, 2011d). Furthermore, the webgraph, social networks, and malaria network were subsequently studied. It is based on the fact that there is structural convergence between these three that we propose to build a search engine which adapts an existing web-search algorithm.

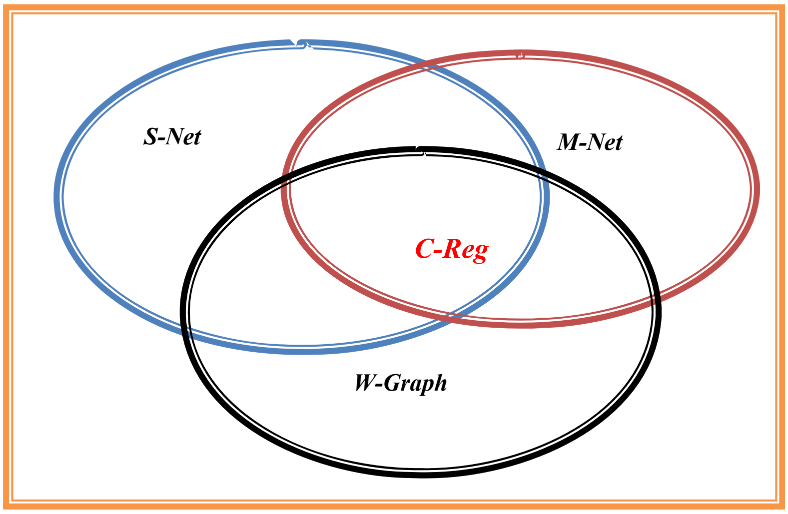
Our task is therefore to examine the three knowledge domains: Social Networks, web graph (a network link of websites), and Malaria Network with the intention of establishing their structural similarities. The overall aim will be to use the result of such comparative analysis to build framework that will give rise to a hybrid search engine for Malaria Networks. The relevance of such a product cannot be overemphasised, since it has the potential to enhance precision-based detection of risky locations for vector control. This implies that results generated from malaria network search could be exploited in such a way that it becomes a form of vector control and avoidance campaign media information, just like the daily weather forecasts that are usually released from meteorological research. Tourists and travellers from around the world could be rightly informed on the level of malaria vector risks expected in such areas, so that adequate precautions could be made before hands.

2 The convergence model

The diagrammatic representation of proposed convergence model is shown in Figure 1, with *S-net*, *M-net* and *W-Graph* representing the Social Networks, Malaria Networks, and web graphs respectively. The area *C-Reg* represents the convergence region. The interest is to discover the attributes and computational methodologies applicable to the region of convergence, *C-Reg*. Major areas of focus are first, the link structure attributes and secondly, the search engine algorithms. After the initial task of discovering the properties of *C-Reg*, it is expected that the *W-Graph* algorithms will be adapted to build a hybrid search engine for *M-net*, while *S-net* algorithms will be applied for necessary performance evaluations (or model validation).

The decision to use social network analysis (SNA) method for evaluation is based on the fact SNA has well developed methods to measure the node importance in the network (Ying et al., 2009). A survey of existing SNA software by Mark et al. (2004) enumerated the following options which are currently used in the industry: UCINET Ver 6.29 (Borgatti et al., 2002), NetMiner Ver 2.4.0 (Cyram, 2004), MultiNet Ver 4.38 (Richards and Seary, 2003) and StOCNET Ver 1.5 (Boer et al., 2004). This is by no means an exhaustive listing of all existing SNA software. While no SNA tool may be described here as infallible, all-encompassing, and superior over others, one thing is certain. And that is the fact that every social network involves contacts between the interacting elements. Since SNA is already in its advanced stage of development, the node contact aspect of such systems could easily be adapted for evaluation of contact networks in disease transmission research. Alternatively, since most of the algorithms that implement the node importance related analysis in social network are in public domain, they could be studied and re-implemented for the malaria contact networks. This way, it would be possible to compare result of web-based search engine with the one from SNA algorithm.

Figure 1 The convergence model (see online version for colours)



3 Framework construction

The web graph is the graph whose nodes are static web pages and edges are directed hyperlinks (Donato et al., 2006). There is an increasing interest in analysing the link structure of the web. For instance, the authors in Donato et al. (2004) reported that they used 300 million nodes web graph provided by Stanford WebBase project for comparative analysis against existing web graph models. As pointed out by Kleinberg and Steve (2001), link analysis has proven useful in the design of improved search engines. Research has it that web link structure is made up of bipartite cores. The occurrence of bipartite graphs in the web was explained by Luciano et al. (2008) in terms of evolving world wide web (WWW), where the authors argued that new nodes are continuously added, connecting to previous nodes with either uniform probability or with preference to older nodes. This behaviour takes places because newly created nodes are likely to be unknown by most page creators, and so have to link to already existing nodes. This is in line with the model of preferential attachment reported by Albert et al. (1999), that at every discrete time step a new vertex is inserted to a number of previous vertices chosen according to the preferential attachment rule (that is probability proportional to the in-degree). The graph of this model shows a power-law distribution over the degree of the vertices. Zhao et al. (2010) mentioned that "*Bipartite graphs are widely used to model social networks and web data*".

Though it is undisputable that due to advancements, web is also made up of objects (e.g., movies, organisations, persons, etc), research has shown that these objects are also accessed via web links. Hence this argument does not invalidate the application of bipartite structure in web models. For instance Liu et al. (2008) applied bipartite graphs for video topic discovery and tracking. In their words, "*The bipartite graph represents the correlation between web videos and their keywords, and automatic topic discovery is achieved through two steps – coarse topic filtering and fine topic re-ranking*". Real life implementation of this was reported to have been successfully done on 'YouKu', a YouTube counterpart in China.

Bipartite model of the web has been applied to search for fraudsters in the web. A research in this category is Furnkranz et al. (2006) in which the authors stated, "*we can clearly observe the existence of a bipartite core between the fraudsters and accomplices*". In a similar way, Grujic et al. (2009) also used bipartite graph to model interactions between users of web databases, where users post comments related to specific subjects. Chen (2010) has described how "*query-concept bipartite graph*" was used to model relevance of queries issued on the web.

3.1 The web graph and social network

A number of researches have established that there is strong similarity between web graph and social networks. The researcher Osinski (2003) is quoted as saying, "*Like social networks in the real world, the Web is clustered into communities, which are groups of Web pages closely related to the same subject*". In the same way, Bing (2011) reiterated that the two popular web search engine algorithms PageRank and HITS originated from SNA. One of the recent researches in support of convergence between social networks and web graphs in terms of analytical methodologies is Xu et al. (2010). Among other methods, the literature discussed the concept of centrality, which is the measure of node influence in networks. The author holds that these analytical measures can be applied in both social networks and web mining. In the words of the author: "*The World Wide Web can be thought of as a virtual society or a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. Therefore, results from social network analysis can be adapted and extended for use in the Web context. In fact, the two most influential link analysis methods, PageRank and HITS, are based on the ideas borrowed from social network analysis*".

We briefly mention here that the research by Broder et al. (2000) which analysed 200 million of web links reported that holistically, the web is shaped like a bowtie. The bowtie shape model grouped the web into major compartments: the *IN*, *CORE*, *OUT* and *DENDRITES* plus the unconnected components. Though details of this shape is not discussed here, we believe that this model does not invalidate, but rather complements the bipartite nature of web, as a bowtie can actually be viewed as bipartite structure with two arms *IN* and *OUT* pointing to *CORE*.

A research has attempted to build geographic search engine based on convergence between social network and web. In this work by Baeza-Yates et al. (2009), titled "*The Geographical Life of Search*", the authors testified that, "*Web is a cultural mirror of society and shed light on the implicit social network behind search*". The researchers studied the relationship between two geographic locations - where user queries are issued and where the web page is hosted.

Another research that underlines the strong relationship between web and social network is Zhang et al. (2010), which maintained that "*communities form a significant structure of the web*". They went further to define a web community as "*a set of web pages concerning a group of individuals sharing a common interest*", pointing out that web communities are "*characterized by dense bipartite subgraphs*", and that finding communities is of "*great help to modern search engines*". The relationship between web and social network can also be seen in the work of Reddy and Kitsuregawa (2001), where authors reported that they can "*mathematically abstract a web community as a dense bipartite graph*", and show that "*community forming is one of the important activities in the Web*".

3.2 The web graph and malaria network

A closer look at the relationship between web graph and malaria network will be taken here. Even though application of contact network modelling in malaria research is to the best of our knowledge relatively new, a number of findings have emerged, which will be mentioned. Research has shown that bites by malaria vectors not only take place in residential homes, but in public places (Eze et al., 2011a), and that if focus on vector control is limited to residential homes, eradication may not be achieved easily. Our research model focuses on human beings with some key attributes, one of such being that their residential homes are under reliable vector control. Another is that they must have suffered from malaria, or are currently suffering from it. Given that the residential homes are under reliable vector controls, the contacts (blood-sucking bites) between human beings and malaria vectors, which most likely happened in the public places, can therefore be utilised to build malaria network. The nodes of the resulting contact network are public places and human beings with malaria history. Mosquito flight behaviour such as flight direction and effects of attractants among others are important for structural definition and modelling of malaria networks. The work done in this area (Eze et al., 2011b) will be outlined in the relevant section of this paper. Previous research has shown that the malaria network is heterogeneous and bipartite (Eze et al., 2011c), and there are algorithms for contact link generation as well as estimation of contact weights (Eze et al., 2011d). Further explanations will also be given in the next section of this paper. One obvious convergence is that both Malaria network and web graph can be represented as bipartite network links. Moreover, every bipartite network can be transformed into a link matrix L, and every link matrix of a bipartite network can also be transformed into an adjacency matrix A as clearly illustrated in equations (1) and (2).

$$L(k,j) = \begin{cases} 1 \; if \;\; X_k \;\; and \; Y_j \;\; are \; linked \\ 0 \; if \;\; X_k \;\; and \; Y_j \;\; not \; linked \end{cases} \tag{1}$$

$$A = \begin{bmatrix} 0 & L \\ L^T & 0 \end{bmatrix} \tag{2}$$

where L^T is the transpose of matrix L.

Given a 2-mode network of nodes $X = \{X1, X2, X3\}$ and $Y = \{Y1, Y2\}$ with cardinalities $N = 3$ and $M = 2$ respectively. Figure 2 gives a pictorial illustration of transformation from graph level into adjacency matrix format.

The popular PageRank algorithm which powers the Google search engine originated from the work of Brin and Page (1998), and ranks the importance of any web page based on the pages that link to it. Hence it looks at web hyperlinks as 'trust votes' and ranks search results based on these links. The PageRank Algorithm has its foundation based on the theory of dominant eigenvalues and corresponding eigenvectors. This has been illustrated in Larry (2009). Apart from PageRank, there are other web search algorithms, for instance the hypertext induced topic search (HITS) algorithm by Kleinberg (1999), and stochastic approach for link structure analysis (SALSA) by Lempel and Moran (2001).

As stated by Ayam et al. (2006), ranking in these three algorithms are all based on derivation of dominant eigenvector of some matrix describing the network, though the way the matrix is defined differs in each case. The link and weight matrices of our model network are both real valued. Furthermore, as already demonstrated in equations (1) and (2), formation of adjacency matrix is also guaranteed. A mathematical transformation can also be utilised to derive a symmetric format for further search engine operations. For instance, as already mentioned by Chandranna (2010), the web search algorithm like HITS generates authority and hub matrices A and H by transformation of the weight matrix W using the following mathematical rules:

$$A = W^T W \tag{3}$$

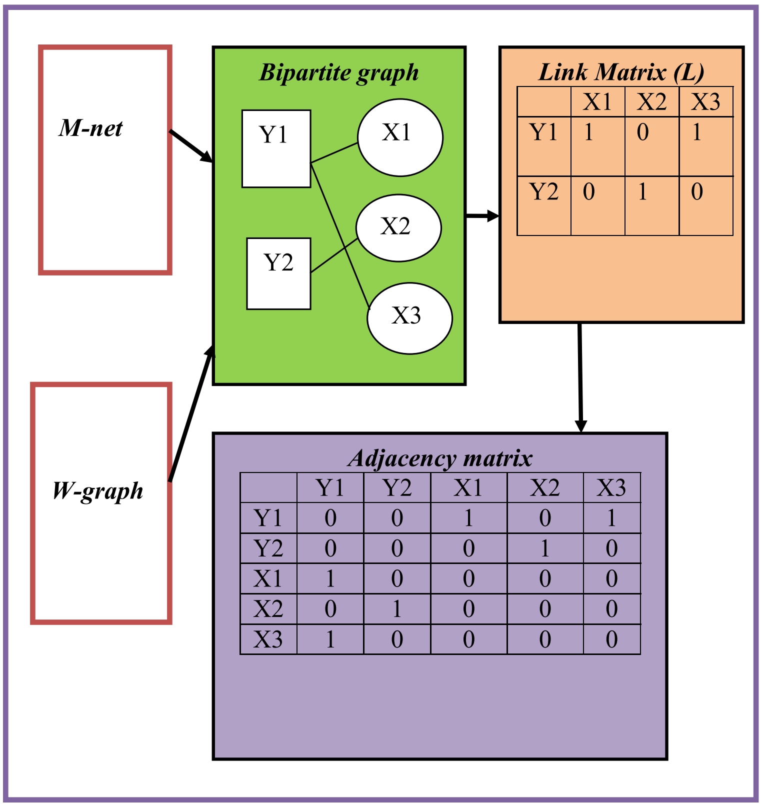
$$H = WW^T \tag{4}$$

where T is the transpose function.

This implies a computational guarantee for a set of real non-trivial eigenvalues and eigenvectors, based on the Perron-Frobenius theory of non-negative symmetric matrices (Richard and Dragos, 2009). An alternative theorem presented by Saad (2011) based on spectral radius is, "*Let A be a real n × n nonnegative irreducible matrix. Then ρ(A), the spectral radius of A, is same as a simple eigenvalue λ of A. There also exists an eigenvector u with positive elements associated with this eigenvalue*".

The overall implication is that the proposed application of web search algorithms on Malaria network is mathematically feasible.

Figure 2 Transformation into adjacency matrix (see online version for colours)



	X1	X2	X3
Y1	1	0	1
Y2	0	1	0

	Y1	Y2	X1	X2	X3
Y1	0	0	1	0	1
Y2	0	0	0	1	0
X1	1	0	0	0	0
X2	0	1	0	0	0
X3	1	0	0	0	0

4 Related works

Our research appears to be a pioneering work in the area of application of web search engine algorithm on contact network, to rank public places based on malaria vector densities. Literatures however indicate that web search algorithms have been successfully used to solve several problems outside the web domains. Here we outline ten major scientific works in this category.

In a research titled "*A Google-like model of road network dynamics and its application to regulation and control*", Crisostomi et al. (2010) applied Google's PageRank algorithm in road network modelling. The authors listed two of the search objectives tackled as discovery of 'traffic congestion' and 'emissions' on the roads.

Web search algorithms have also been applied in biological networks. A very recent research in this category is Ivan and Grolmusz (2011) which focused on protein networks. The authors pointed out that they adapted their solution technique from PageRank algorithm, and that the overall model is "*fast, scalable and robust, and its capabilities are demonstrated on metabolic network data of the tuberculosis bacterium*".

In the field of genetic engineering, web search algorithms have also been applied in gene prioritisation. Chen et al. (2009) in their research mentioned two algorithms called '*PageRank with Priors*' and '*HITS with Priors*', which are said to be adaptations of Google's PageRank and Jon Kleinberg's HITS web search algorithms respectively. In the words of the authors, the research demonstrates that "*methods used for studying both social and Web networks can be successfully used for disease candidate gene prioritization*".

Lexical analysis problems have been solved using web search algorithms. An instance of this is the work of Navigli and Lapata (2010) which dealt with word sense disambiguation. This research first builds a graph from a lexical knowledge base of all possible interpretations of the word sequences. The authors explained that graph nodes and edges correspond to word senses and word dependencies respectively. They further gave instances of dependencies as '*synonym*' and '*antonym*'. The search algorithm is then used to carry our sense disambiguation, a process which was explained as "finding the most important node".

Another scientific area that has applied web search algorithm to solve non-web problems is the field of ecology. Allesina and Pascual (2009) explained the issue of forecasting species' extinction, which the authors described as "*a major challenge in ecology*". In explaining how web search algorithm was adapted to solve this problem, the authors said that "*an algorithm adapted from the one Google uses to rank web-pages can order species according to their importance for co-extinctions, providing the sequence of losses that results in the fastest collapse of the network*". While evaluating the project performance, the authors reiterated that "*the algorithm finds the best possible solution for the problem of assigning importance from the perspective of secondary extinctions in all analyzed networks*". They further explained that their approach relies on network structure.

Yu et al. (2010) described how bipartite models can be applied in a dynamic manner. Two of such cases the authors dealt with are 'user-movie bipartite graph', where the links represent movie ratings given by users, and the 'author-conference bipartite graph' where the links are the number of papers published by the corresponding author.

Other projects in this category are features selection in data mining (Dino et al., 2008), ranking authors in co-citation networks (Ying et al., 2009), Ranking the importance of boards of directors (Martin and Mark, 2004), etc. Furthermore, Zhou et al. (2012) proposed a model which applies web search algorithm on '*author–paper bipartite networks*', with the aim of evolving appropriate metric for evaluating the scientific credit of papers, scientists and journals. The researchers used this work to support their argument, that the ranking assigned to scientific publications should not be a simple function of number of citations, but rather a metric which puts into consideration the prestige of scientists, as well as the quality of their publications.

Like the emergence of search engines revolutionised web research, and led to applications in diverse scientific fields enumerated above, it is expected that application of search engines in malaria research will make a tremendous impact in malaria eradication. A breakthrough in this area is expected to advance the field of malaria research, and also provide strong reference base for other active researchers in network computing, web mining, pattern analysis, etc. since all these fields seem to have common traits.

5 Supporting experiments/previous works

Some of the experiments already undertaken at a more fundamental level of this research will be outlined in this section. Appropriate references will be provided in cases where they have been published in peer-reviewed journals or conference proceedings.

5.1 Vector existence computational research

At the conceptual level of building contact network models for research on malaria transmission in public places, one of the questions tackled is the issue of vector existence in public places. Effort was made to computationally determine if vector existence in public places is really a universal concept, or just a localised one. The mini project done in this regard is the *vector existence research* by Eze et al. (2011a). This is a survey research carried out in the Sarawak State of Malaysia in June 2010 on vector existence in public places. The computational results obtained from this experiment led to many scientific discoveries, and also underlined the necessity for application of search engine in contact network modelling of malaria transmission in public places. Appropriate questionnaire was designed, and a total of 525 valid responses were gotten from respondents drawn from ten countries as shown in Table 1.

Table 1 Country of origin of respondents

Country	*Respondents count*
China	1
Germany	1
Indian	3
Indonesia	5
Korea	1
Libya	1
Nigeria	3
Pakistan	1
Palestine	2
Malaysia	507
TOTAL	525

Source: Eze et al. (2011a)

The foreigners involved are those who visited, or were living in Malaysia at the time of this research. The aim of this survey is to determine previous experiences of mosquito bites in 16 carefully selected types of public places listed in Table 2. The third column of the table represents the count of number of respondents who have experienced mosquito bites in the corresponding public places (in second column). Respondents were also required to select three of the public places where they experienced highest degree of mosquito bites, and to assign ranks to those places.

Table 2 Vector existence count table

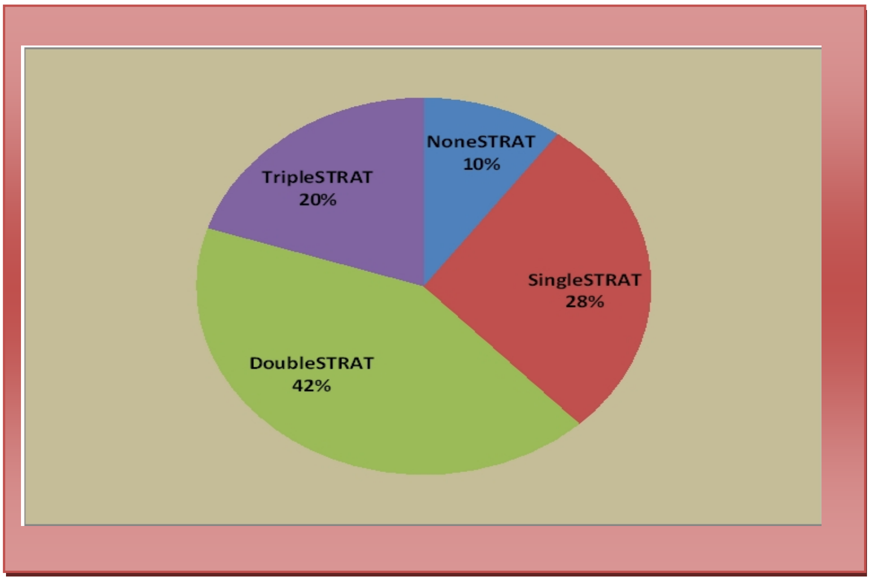
S/N	*Public place*	*Respondents count*
1	School	398
2	Foot ball pitch	156
3	Bus stop	317
4	Market	247
5	Swim pool	86
6	Restaurant	207
7	Airport	30
8	Road	226
9	Barb saloon	26
10	River	442
11	Zoo	249
12	Mountain	283
13	Hotel	76
14	Office	60
15	Cinema hall	50
16	Night clubs	26

Source: Eze et al. (2011a)

Based on the outcome of this experiment, as much as 95% of the respondents use vector control tools in their residential homes, while only 5% live without vector controls. The strategies employed by the respondents are: use of *bed nets*, *insecticides* and *window nets*. Computational result showed that notwithstanding the fact that as much as 95% of the respondents used vector control in their residential homes, many of them still suffered from malaria. The pertinent question is, "Why is this anomaly?" The quest for answer to this question led to a detailed vector control combination analysis, for the respondents who suffer from malaria.

The outcome is shown in Figure 3, with four codes designed to represent the combination of vector control strategies used by respondents. These are *NoneSTRAT*, *SingleSTRAT*, *DoubleSTRAT*, and *TripleSTRAT* for none, single, double and triple strategies respectively. This implies that as much as 62% of those who suffer from the disease use at least two vector control strategy. Since it is assumed that the respondents are not naïve in the use of vector control tools based on other attributes like level of education, age, etc which were studied, there is one way to account for why they still suffered from malaria even after applying vector controls in their residential homes. That is the fact that they most likely contacted the malaria from the public places rather than in their residential homes.

Figure 3 Vector strategies combination chart (see online version for colours)

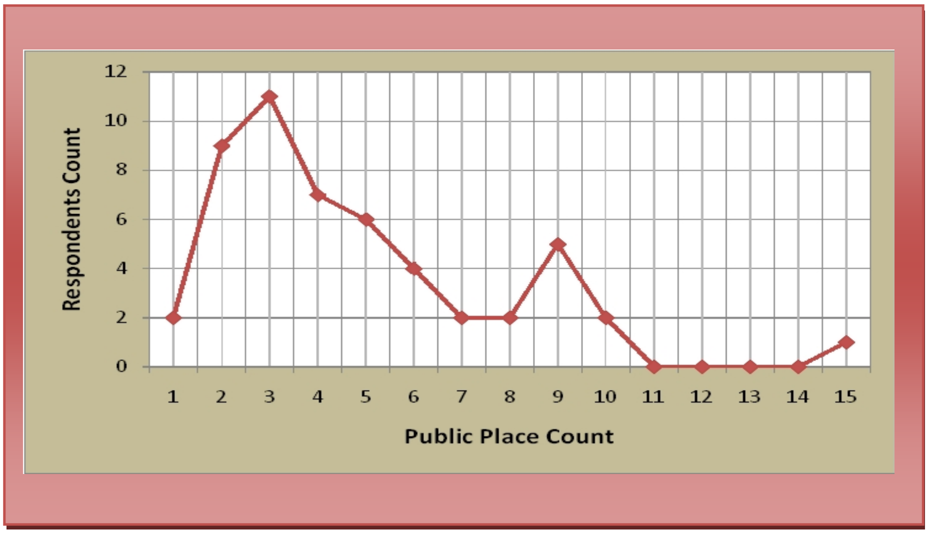


Source: Eze et al. (2011a)

Computational analysis was also done to discover cases of mosquito bites with infection, as well as bites without infection. The graphical outputs of these analyses are shown in Figures 4 and 5 respectively. We will summarise findings as it relates to search engine. The vector bites with infection graph relates to respondents who were bitten by mosquitoes in public places, and were infected with malaria. On the other hands, the vector bites without infection graph relates to respondents who were bitten by mosquitoes in public places, but were not infected with malaria. Both graphs are plots of the respondents count against the number of public places count.

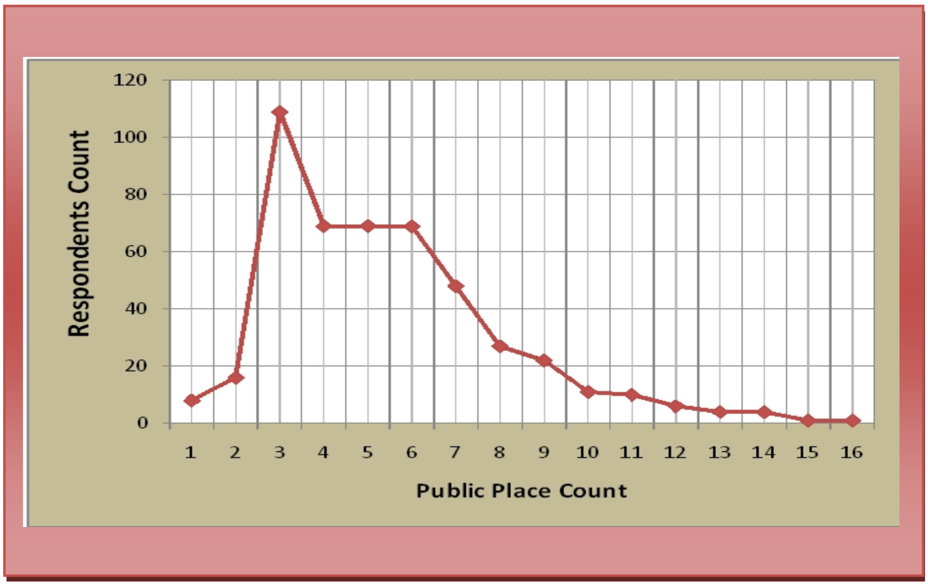
First, this finding confirms the fact already established by scientific researches such as WHO (2009), Donovan et al. (2007) and International Association for Medical Assistance to Travellers IAMAT (2011) that not all mosquito bites lead to malaria, but those of malaria infected vectors. It also underlines the need to focus interest on building models that will locate vector infected public places, rather than just any area populated by mosquitoes. Furthermore, a closer look at the shapes of the two graphs reveals a fairly common trend. For instance, both graphs rise up steadily, reach a peak, and then descend. Many facts can be deduced from this trend. The most important is that there must be a public place or a combination of public places that account for the highest vector density; otherwise there would not have been any peak point as observed. This implies that a search engine is necessary to determine the exact public places concerned.

Figure 4 Vector bite with infection graph (see online version for colours)



Source: Eze et al. (2011a)

Figure 5 Vector bite without infection graph (see online version for colours)



Source: Eze et al. (2011a)

Although the full details of ranking analysis and validation of this survey exist in Eze et al. (2011a), it is important to summarise here that from the outcome of this experiment, there is vector existence in public places. For instance, all 525 respondents who participated in this research mentioned at least one public place where they have had mosquito bites. Moreover, up to two respondents indicated that they have experienced mosquito bites in all the 16 types of public places under investigation. It can therefore be deduced that the concept of vector existence in public places is not limited to Malaysia alone, since respondents were drawn from ten countries. Result of this experiment also clearly proves that even if vectors are effectively controlled in residential home, malaria will still persist as long as the public places still harbour malaria vectors. The aim of a search engine is to locate the public places that habour malaria vectors, and rank them in terms of vector density.

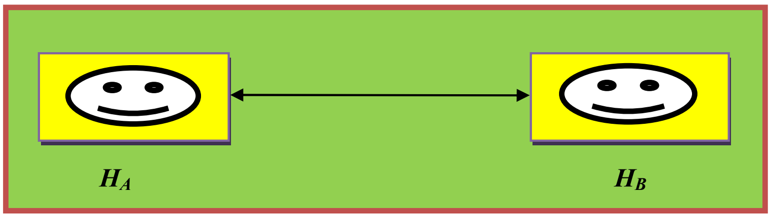
5.2 *Heterogeneous contact network evolutionary research*

It is natural for every search engine to have a search space, which is like a parameter on which it operates. Searches therefore take place with the aim to locate the item, or group of items, which meet desired criteria. It is in line with this analogy that prior research efforts were invested in the area of evolving malaria network for the proposed search engine. The outcome of those previous experiments will be outlined here.

5.2.1 *The binary tree-based heterogeneous network link model*

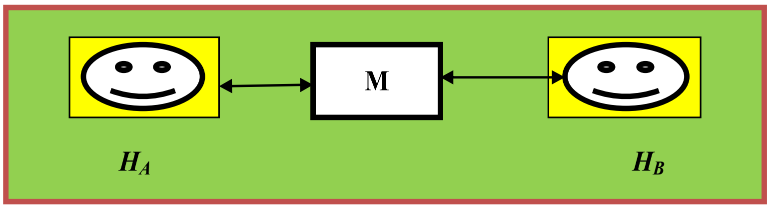
A research titled "The Binary Tree-Based Heterogeneous Network Link Model" by Eze et al. (2011c) dealt with evolution of contact network for malaria transmission study. Among other things, this research tackled the issue of determining the structure of the network, and then evolved a link algorithm that builds the desired heterogeneous network. The nodes of the network are the public places and human beings involved in malaria transmission. The research did a comparative analysis of link representations for set of human beings, e.g., $H = \{H_A, H_B\}$, and public places (habouring malaria vectors M) in contact networks as shown in Figures 6 and 7. The work further proved that Figure 6 is faulty, while Figure 7 is the ideal format in terms of contact links representation.

Figure 6 Node link representation – faulty version (see online version for colours)



Source: Eze et al. (2011c)

Figure 7 Node link representation – ideal version (see online version for colours)



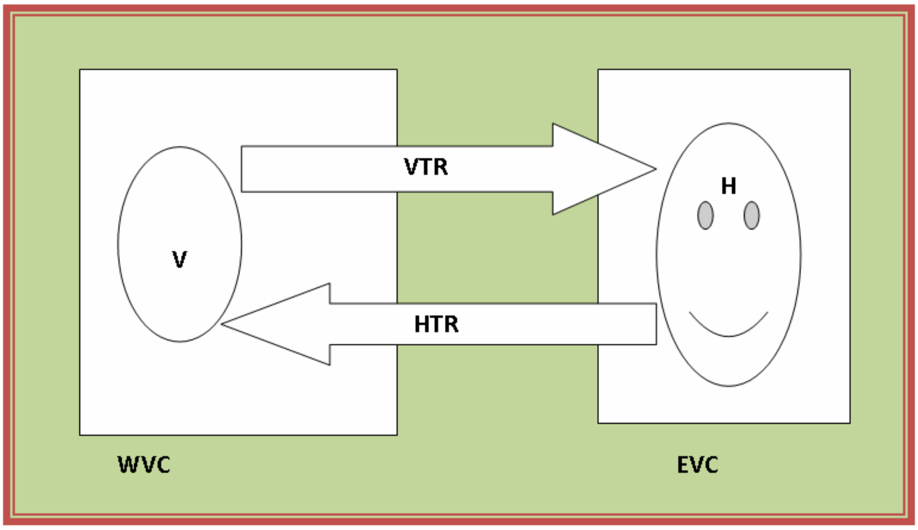
Source: Eze et al. (2011c)

It was further shown that the network structure is bipartite, and that a homogeneous network is not sufficient for use in building a malaria network. Further steps were also taken to adapt the general proof of bipartition in Akiyama and Kano (2011) to show that malaria network is bipartite. Implementation of the binary tree generating algorithm was done in MATLAB.

5.2.2 Node contact strength generating algorithm

After building the contact network at the link level, the next task undertaken is the issue of determining the strength or relationship between the heterogeneous nodes in a malaria network. In simple terms, the task involves modelling the degree of relationship (edge weights) between the human beings and the public places nodes in a contact network. This is because human beings are only expected to get infected with the disease by being in contact with infected vectors in public places, and the more the degree of attachment, the more the possibility of infection. The contact strength generating algorithm evolved by Eze et al. (2011d), which has been presented in a peer-reviewed international conference tackles this issue. Some highlights of this algorithm will be given here. First, the research analysed the malaria transmission environment, beginning from the malaria lifecycle. The transmission environment is then conceptualised. Figure 8 shows a typical transmission environment adapted from by Eze et al. (2011d), where *V*, *H*, *VTR* and *HTR* represent the vectors, human beings, vector traffic (related to vectors flight behaviours) and human traffic respectively. Furthermore, *WVC* represents the environment without vector control, while *EVC* represents the environment with vector control.

Figure 8 Simulation environment (see online version for colours)



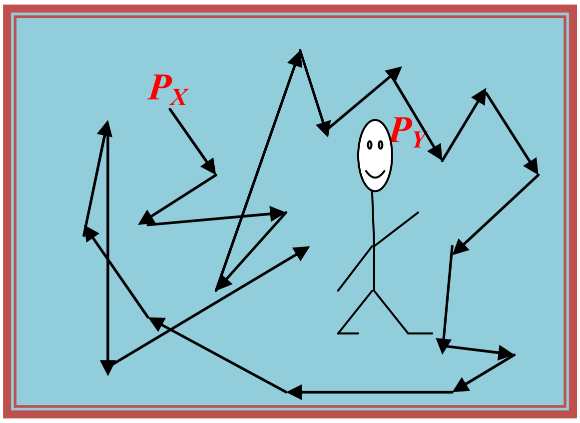
Source: Eze et al. (2011d)

These and many other parameters were put into consideration in evolving the contact strength generating algorithm. Hence the contact strength algorithm is used to define the edge weight of the bipartite network originally generated by the link algorithm (mentioned in 5.2.1).

5.2.3 The mosquito flight behaviours

Further studies were done on the flight behaviour of malaria vectors, since this was found to be important in contact network model development. This study was based on literature survey. The outcome of this study (Eze et al., 2011b) shows that a number of things influence the flight behaviour of malaria vectors. For instance, carbon dioxide and human odour (Carrie, 2005) among other things attract the vectors, and also influence their flight behaviours. A number of statistics such as maximum flight capability were considered. In summary, literature shows that malaria vectors undertake flight pattern that is zigzag (Rafael and Ricardo, 2009), and influenced by attractants (Gu et al., 2006). Figure 9 shows a typical flight from arbitrary position P_X towards a human being in position P_Y (Eze et al., 2011b), under the influence of attractants.

Figure 9 Mosquito in zigzag flight, attracted to human host (see online version for colours)



Source: Eze et al. (2011b)

5.2.4 Web link structure experiment

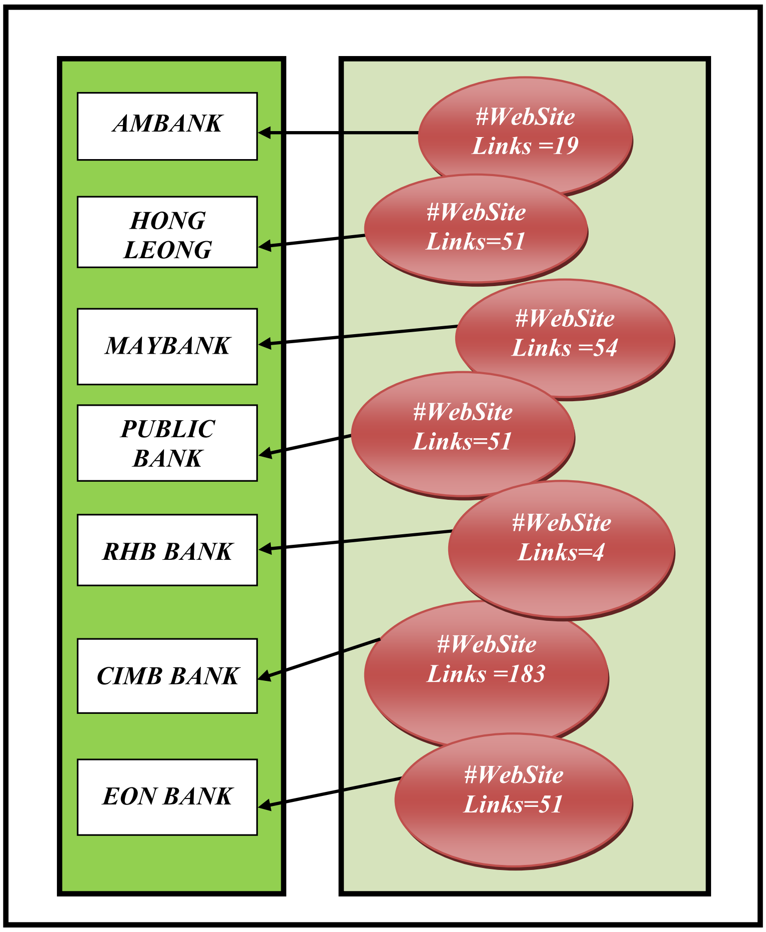
One of the experiments carried out to study the bipartite link structure of web graph will be described in this section. The dataset consisted of web links mined from the websites of selected service companies through a web crawl operation. The choice of these institutions was based entirely on the need to use well established organisations with similar product and service interest as case study. Selecting the nodes with similar product and service attribute is in line with the definition of bipartite network as made up of two major partitions, each partition having its elements different from those of the other partition. Table 3 shows the names and websites of the seven companies selected among the major banking institutions in Malaysia for this study.

To obtain the dataset for the study, a web crawler WinWeb 2.0 (WinWeb, 2011) was used to mine the websites that are pointing (hyperlinked) to these seven companies. The initial result of the data extraction process is shown in Figure 10. The right hand (RH) side of this figure consists of seven groups of websites, pointing to the seven companies. The set of number of web links involved at time of this experiment is {19, 51, 54, 51, 4, 183, 51} as indicated. This structure no doubt closely compares with the structural link between human beings and the public places in a malaria transmission network. Careful examination shows that the web link structure in Figure 10 can be neatly superimposed on a malaria network of seven public places and 413 human beings.

Table 3 DataSet source table

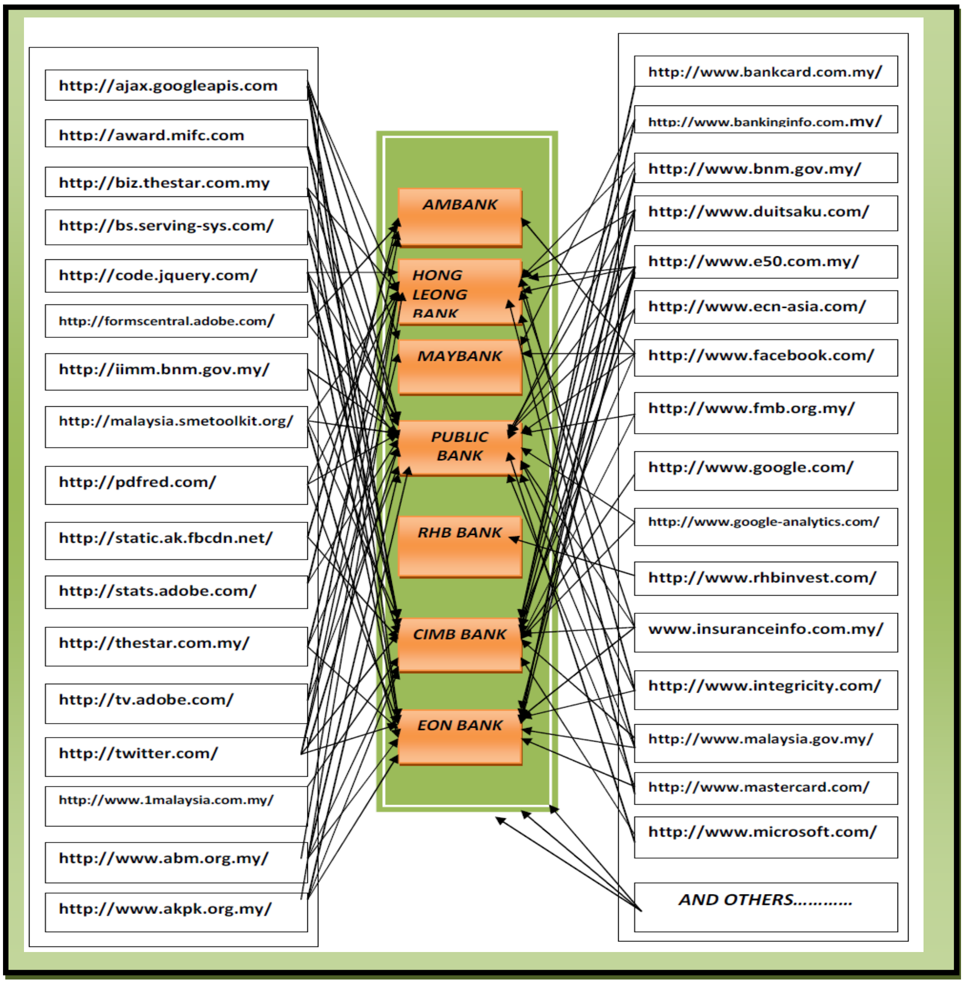
S/N	*COMPANY*	*WEBSITE*
1	AmBank	http://www.ambg.com.my
2	Hong Leong Bank	http://www.hlb.com.my
3	Maybank	http://www.maybank.com
4	Public Bank	http://www.publicbank.com.my
5	RHB Bank	http://www.rhb.com.my
6	CIMB Bank	http://www.cimb.com
7	EON Bank	http://www.eonbank.com.my

Figure 10 Core nodes linked to community of websites (see online version for colours)



This structural resemblance can be clarified by looking at the left hand (LH) structure as seven public places, rather than seven banks, and the RH structure as 413 human beings, rather than 413 websites. Further analytical steps taken to break the dataset into individual web links gave rise to Figure 11 which is approximately bipartite in structure. To further substantiate this result, the analysis revealed that none of the seven core nodes pointed to one another, though this is not always guaranteed for every website, since the decision to link or not to link is entirely dependent on website creators. We use the term '*approximately bipartite*' because this experiment fully confirmed that the seven companies do not point to themselves, but did not perform full analysis for all the websites that point to them – a step which was assumed to be unnecessary. These remaining unconfirmed websites were simply grouped as 'OTHERS' as shown in Figure 11.

Figure 11 Detailed bipartite link structure (see online version for colours)



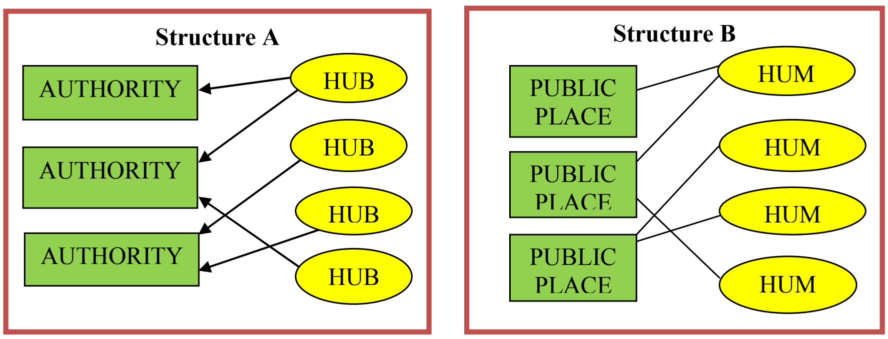
6 Proposed methodology

Two major web search algorithms, PageRank and HITS were reviewed for use in building the proposed search engine for malaria transmission research. The later was chosen, and the reason for this choice will be explained in this section.

6.1 Choice of search algorithm

The acronym HITS stands for hypertext induced topic search (Chandranna, 2010), a web page ranking algorithm. The Alta Vista Search engine was implemented based on HITS (Debajyoti et al., 2006). A research by Ravi et al. (2006) also pointed out that the Clever System of IBM Almaden Research Center used the HITS Algorithm. Unlike PageRank which is static and query independent, the HITS is a query-dependent algorithm (Mikhail and Ryen, 2008). The fact that the proposed malaria search engine uses a malaria network of specific dimension, with well defined strength of relationship between the heterogeneous nodes (analogous to query dependence concept), is one obvious reason for choice of HITS algorithm. Furthermore, HITS algorithm looks at the entire search space as made up of 'authority/hub' structure (Qing-Xian and Li, 2011), a format which tightly fits into proposed malaria search engine which looks at the malaria network search space as a bipartite network of 'public places/human beings' structure. This appears to be like a one-to-one mapping between 'authority/hub' and 'public places/human beings' when the two structures are comparatively examined as depicted in Figure 12.

Figure 12 The 'authority/hub' *versus* 'public places/human beings' diagram (see online version for colours)

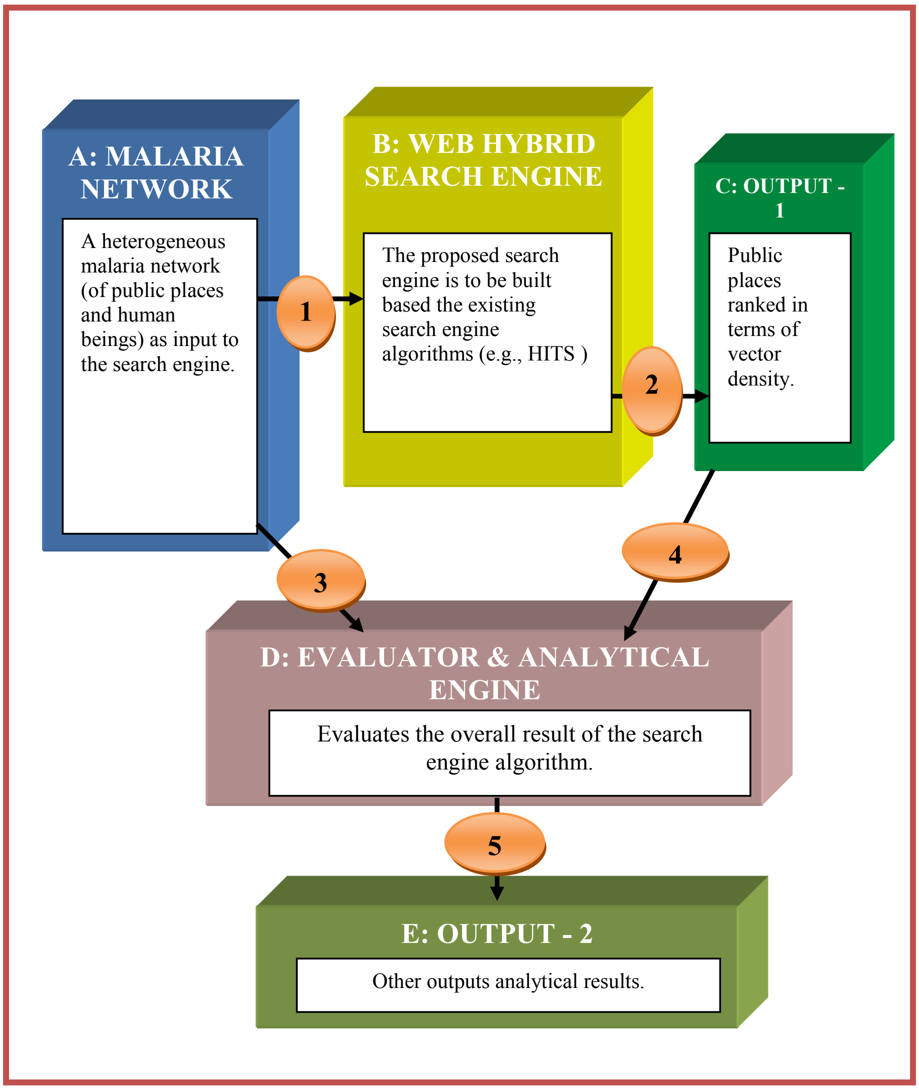


6.2 Summary of workflow for proposed model

The workflow for the proposed project is shown in Figure 13. The core module which performs the search is shown as block *B*. As indicated in the diagram, the search engine is an adaptation (hybrid) of existing web-search algorithms (e.g., *HITS)*. The input block *A* is the Malaria Network, which has previously been described in the literature review section. The initial output of the search engine is indicated as block *C*, and consists of the ranking of public places using the search engine. This output ranks the public places in terms of calculated level of infective malaria vector densities. Various output formats such as sorting in order of public places, or in order of calculated densities are proposed.

Hence the public places of utmost concern, which accounts for highest infection of the human beings under investigation, can easily be spotted by this initial search output. Based on the 'authority/hub' structure of HITS algorithm, this approach can also spot human beings whose activities lead to greatest infection in the network. We mention this because though our search engine is focused on location of public places, HITS algorithm can search both in terms of hubs or authority.

Figure 13 The block diagram of the proposed framework (see online version for colours)



Block *D* which is the evaluation and analytical module performs various forms of evaluation and analysis. As shown in the diagram, the input arrows '*3*' and '*4*' emanate from blocks *A* and *C*. This section evaluates the overall performance and results of the search operations. A number of other algorithms tailored towards mining the Malaria

Network are also proposed for this section, one of which is the measure of human crowding in the public places. Another is the question of indirect transfer of malaria from one public place to another. For instance, one could be interested in discovering how a serious infection on a given public place will affect another public place, as well as what measure of positive effect will result in other public places if there is a successful control in a particular public place. Finally, various outputs from the proposed evaluations and analysis will form the second output section designated as block *E*.

7 Discussion on proposed model

In this paper, we proposed how structural convergence between web graph, social network and malaria network can be leveraged to build a search engine for discovering public places that habour malaria vectors. Some of the relevant characteristics of the web graph, malaria network and social networks were outlined. We have mentioned several researches that solved problems outside web domains, using web search algorithms. Two major structural similarities were covered. First is the similarity in terms of bipartite structure. The second is more specific, and is related to the proposed web search engine (HITS), where the 'authority/hub' structure is closely compared with the 'public places/human beings' structure in the malaria network. Several schematic diagrams have also been presented, including the summary of the proposed workflow. Furthermore, some of the earlier experiments performed in this research were also outlined, and some relevant outcomes presented.

Referring back to the convergence model diagram in Figure 1, the region of convergence depicted as *C-Reg* represents the structural resemblance between web graph, social network and malaria network as explained within this paper. It is based on the existence of such convergence that a search engine is built as an adaptation (or hybrid) of existing web-algorithm search engine, and tailored for use in malaria networks. It is also based on this that we proposed to use a SNA system for search engine validation.

8 Conclusions

A lot of literature surveys have been reviewed on application of web search algorithms in solving diverse problems in non-web domains. Conceptual models have been built from previous researches in the area of malaria transmission, including the evolution of malaria network, as has been pointed out in this paper. Several experiments have also been performed, some of which were summarised in this paper.

The major aim of this paper is to build a framework in order to exploit the structural similarities between social networks, web graph and malaria network for advancement in malaria research through search engine operations. The fact that these three knowledge domains are rooted in graph theory underscores the possibility of opening a new frontier in malaria research through application of theories and methods of web graph and social networks. It is expected that some of the problems associated with real-life web search engines (e.g., spamming) will not be experienced in malaria network search. Spamming which is an unethical computing technique was explained by Christopher et al. (2009) as the act of pointing misleading web links to a website to boost its ranking. Detecting and combating such systematic abuse is an important task in web search engine design. Due to strict control measures in malaria network, spamming is not anticipated.

In conclusion, the stage appears to be set for scientists to build web-hybrid search engine for malaria research by exploiting the structural similarities between social networks, web graph and malaria network. This resulting system will no doubt lead to advancement in malaria research.

9 Future work

This research is in progress, and the next stage is the implementation of the proposed system, both as a prototype and a real live evaluated system. Apart from the proposed search engine, it is believed that a successful conceptualisation of the malaria network so far achieved in the previous experiments has set the stage for further analytical research in this field. Development of new visualisation algorithms tailored towards showing the outputs of various analytical results is also being pursued vigorously as part of future work.

Acknowledgements

The authors acknowledge with gratitude the Ministry of Higher Education for financial support through the Fundamental research grant scheme FRGS/2/10/SG/UNIMAS/02/04.

References

Abirami, D., Shobana, K. and Murugan, K. (2011) 'Larvicidal, pupicidal and smoke toxicity effect of Kaempferia Galanga to the malaria vector, Anopheles Stephensi', *The Bioscan. J.*, Vol. 6, No. 2, pp.329–333.

Akiyama, J. and Kano, M. (2011) *Factors and Factorizations of Graphs: Proof Techniques in Factor Theory*, Springer-Verlag, Berlin Heidelberg.

Albert, R., Jeong, H. and Barabasi, A. (1999) 'Diameter of the world wide web', *Nature*, No. 401, p.130.

Allesina, S. and Pascual, M. (2009) 'Googling food webs: can an eigenvector measure species' importance for co-extinctions?', *PLoS Comput. Biol.*, Vol. 5, No. 9, e1000494, doi:10.1371/journal.pcbi.1000494.

AMCA (American Mosquito Control Association) (2011) *Mistaken Identities: Insects Commonly Mistaken for Mosquitoes*, American Mosquito Control Association, Mount Laurel, NJ 08054 [online] http://www.mosquito.org/mistaken-identities (accessed 5 October 2011).

Ayam, F., Thomas, L., Joel, M., Gregory, R. and Lesley, W. (2006) 'Authority rankings for HITS, PAGERANK, AND SALSA: existence, uniqueness, and effect of initialization, society for industrial and applied mathematics (SIAM)', *J. SCI. COMPUT.*, Vol. 27, No. 4, pp.1181–1201.

Baeza-Yates, R., Middleton, C. and Castillo, C. (2009) 'The geographical life of search', *Proceedings of IEEE/WIC/ACM International Joint Conferences WI-IAT '09*, 15–19 September, pp.252–259.

Bing, L. (2011) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, New York.

Boer, P., De Negro, R., Huisman, M., Snijders, T.A.B., Steglich, C.E.G. and Zeggelink, E.P.H. (2004) *STOCNET: An Open Software System for the Advanced Statistical Analysis of Social Networks*, Ver. 1.5., University of Groningen, Broerstraat, Groningen, Netherlands.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002) *UCINET 6 for Windows: Software for Social Network Analysis*, Analytic Technologies, Harvard.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and ISDN Systems*, Vol. 30, Nos. 1–7, pp.107–117, doi:10.1016/s0169-7552(98)00110-x.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, A. (2000) 'Graph structure in the web', *Proceedings of the Ninth International Conference on The World Wide Web*, pp.309–320.

Carrie, M.B. (2005) *Effect of Weather on Mosquito Biology, Behaviour and Potential for West Nile Virus Transmission on the Southern High Plain of Texas*, PhD dissertation in Environmental Toxology, Texas Tech University, Texas, USA.

Chandranna, A.K. (2010) *Implementing an online Version of Hyperlink-Induced Topic Search (HITS) Algorithm: A CS 297 Project Report*, May, MSc thesis in the Department of Computer Science, San Jose State University, San Jose, CA, USA.

Charoenpanyanet, A. and Chen, A. (2008) 'Satellite-based modelling of anopheles mosquito densities on heterogeneous land cover in Western Thailand', *The Int. Archives of the Photogrammetry, Remote Sensing & Spa., Info. Sc. J.*, Vol. 37, Part B8, pp.159–164.

Chen, J., Aronow, B.J. and Jegga, A.G. (2009) 'Disease candidate gene identification and prioritization using protein interaction networks', *BMC Bioinformatics*, Vol. 10, No. 1, p.73.

Chen, Y. (2010) *Enhanced Web Search Engines with Query-Concept Bipartite Graphs*, Computer Science Dissertations found in Digital Archive @ Georgia State University, Paper 54.

Christopher, M., Prabhakar, R. and Hinrich, S. (2009) *An Introduction to Information Retrieval*, Cambridge University Press, England.

Crisostomi, E., Kirkland, S. and Shorten, R. (2010) 'A Google-like model of road network dynamics and its application to regulation and control', *International Journal of Control*, 29 July, Vol. 10, No. 53, pp.1–26.

Cyram Co., Ltd. (2004) *Cyram NetMiner II*. Version 2.4.0, Seoul.

Debajyoti, M., Pradipta, B. and Young-Chon, K. (2006) 'A syntactic classification based web page ranking algorithm', *6th International Workshop on MSPT Proceedings*, pp.83–92.

Dino, I., Rosa, M. and Marco, B. (2008) 'Using PageRank in feature selection', *Proc. of the 16th Italian Symposium on Advanced Database Systems (SEBD 2008)*, 22–25 June, pp.93–100.

Donato, D., Laura, L., Leonardi, S. and Millozzi, S. (2004) 'Simulating the WebGraph: a comparative analysis of models', *Computing in Science & Engineering*, Vol. 6, No. 6, pp.84–89.

Donato, D., Laura, L., Leonardi, S. and Millozzi, S. (2006) 'Algorithms and experiments for the webgraph', *Journal of Graph Algorithms and Applications*, Vol. 10, No. 2, pp.219–236.

Donovan, M.J., Messmore, A.S., Scrafford, D.A., Sacks, D.L., Kamhawi, S. and McDowell, M.A. (2007) 'Uninfected mosquito bites confer protection against infection with malaria parasites', *Infection and Immunity J.*, May, Vol. 75, No. 5, pp.2523–2530.

Ethan, A.H. (2010) *Biotech Company Secretly Releases Millions of GM Mosquitoes in Cayman Islands* [online] http://www.naturalnews.com (accessed 19 November 2010).

Eze, M., Labadin, J. and Lim, T. (2011a) 'Emerging computational strategy for eradication of malaria', *The Proceedings of 2011 IEEE Symposium on Computers & Informatics (IEEE/ISCI 2011)*, 20–22 March, pp.715–720, Kuala Lumpur.

Eze, M., Labadin, J. and Lim, T. (2011b) 'Mosquito flight model and applications in malaria control', *Proc. of 3rd International Conference on Computer Engineering and Technology (ICCET 2011)*, 17–19 June, pp.59–64, Kuala Lumpur.

Eze, M., Labadin, J. and Lim, T. (2011c) 'The binary tree-based heterogeneous network link model for malaria research', *Proceedings of 7th International Congress for Industrial and Applied Mathematics Conference (ICIAM 2011)*, 18–22 July, pp.546–547, Vancouver Canada.

Eze, M., Labadin, J. and Lim, T. (2011d) 'Contact strength generating algorithm for application in malaria transmission network', *Proceedings of 7th International Conference on IT in Asia (CITA 2011)*, 11–14 July, pp.21–26, Kuching, Sarawak, Malaysia.

FMLAB (Florida Medical Entomology Laboratory) (2011) *Mosquito Taxonomy & Identification*, 5 October [online] http://fmel.ifas.ufl.edu/research/taxonomy.shtml (accessed 5 October 2011).

Furnkranz, J., Scheffer, T. and Spiliopoulou, M. (Eds.) (2006) 'Detecting fraudulent personalities in networks of online auctioneers', *PKDD 2006*, pp.103–114, LNAI 4213.

Global Risk Forum (2009) *GRF Fact Sheet (2009): Malaria*, p.1, Davos Switzerland [online] [online] http://www.grforum.org/userfiles/file/GRF%20Factsheets/11122009_Factsheet_Malaria_en_final.pdf (accessed 20 February 2010).

Grujic, J., Marija Mitrovic, M. and Tadic, B. (2009) 'Mixing patterns and communities on bipartite graphs on web-based social interactions', *The 16th International Conference on Digital Signal Processing (DSP 2009)*, 5–7 July, Santorini, Greece.

Gu, W., Regens, J., Beier, J. and Novak, R. (2006) 'Source reduction of mosquito larval habitats has unexpected consequences on malaria transmission', *Proc. Natl. Acad. Sci.*, pp.17560–17563, USA.

IAMAT (2011) *How to Protect yourself against Malaria*, International Association for Medical Assistance to Travellers, Ontario Canada [online] http://www.iamat.org/pdf/protect_yourself_against_malaria.pdf (accessed 30 December 2011).

Ivan, G. and Grolmusz, V. (2011) 'When the web meets the cell: using personalized PageRank for analyzing protein interaction networks', *Bio Informatics Original Paper*, 12 December, Vol. 27, No. 3, pp.405–407.

Jonathan, S. (2011) 'Imagining a world without mosquitoes', *The Daily Caller* [online] http://dailycaller.com/2010/09/21/imagining-a-world-without-mosquitoes/ (accessed 3 October 2011).

Kleinberg, J. (1999) 'Authoritative sources in a hyperlinked environment', *Journal of the ACM*, Vol. 46, No. 5, pp.604–632.

Kleinberg, J. and Steve, L. (2001) 'The structure of the web', *Science*, 30 November 30, Vol. 294, No. 5548, pp.1849–1850.

Larry, K. (2009) *Linear Algebra: A First Course with Applications*, Chapman & Hall/CRC Press, New York.

Lempel, R. and Moran, S. (2001) 'SALSA: the stochastic approach for link-structure analysis', *ACM Transactions on Info. Systems*, April, Vol. 19, No. 2, pp.131–160.

Liu, L., Sun, L., Rui, Y., Shi, Y. and Yang, S. (2008) 'Web video topic discovery and tracking via bipartite graph reinforcement model', *WWW 2008*, 21–25 April, pp.1009–1018, Beijing.

Luciano, F., Osvaldo, O., Gonzalo, T., Francisco, A., Paulino, R., Lucas, A., Matheus, P. and Luis, E. (2008) *Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications*, 16 September, pp.35–38, Universidade de Sao Paulo, Brazil & Umea University, Sweden.

Mark, H., Marijtje, A. and Van Duijin, J. (2004) *Software for Statistical Analysis of Social Networks*, University of Groningen, Broerstraat, Groningen, Netherlands.

Martin, C. and Mark, M. (2004) *Ranking the Importance of Boards of Directors*, August, University of Pennsylvania, USA and University of Manchester, UK.

Mikhail, B. and Ryen, W.W. (2008) 'Mining the search trails of surfing crowds: identifying relevant websites from user activity', *WWW 2008*, 21–25 April, Beijing, China.

Naoko, N., Yoshihiko, H., Mutsuo, K. and Akira, I. (2002) 'Analysis of malaria endemic areas on the Indochina Peninsula using remote sensing', *Jpn. J. Infect. Dis.*, Vol. 55, No. 5, pp.160–166.

Navigli, R. and Lapata, M. (2010) 'An experimental study of graph connectivity for unsupervised word sense disambiguation', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 32, No. 4, pp.678–692.

Osinski, S. (2003) *An Algorithm for Clustering of Web Search Results*, June, Masters thesis, Poznań University of Technology, Poland.

Qing-Xian, .W. and Li, L. (2011) 'Map reduce for HITS algorithm with application to Chinese word networks', *Journal of Convergence Information Technology*, September, Vol. 6, No. 9, pp.178–185.

Rafael, M. and Ricardo, L. (2009) 'Presumed unconstrained dispersal of Aedes Aegypti in the city of Rio de Janeiro, Brazil', *Rev Saudi Publica Journal*, Vol. 43, No. 1, pp.8–12.

Ravi, K., Prabhakar, R. and Andrew, T. (2006) 'Core algorithms in the CLEVER system', *ACM Trans. on Internet Technology*, May, Vol. 6, No. 2, pp.131–152.

Reddy, K.P. and Kitsuregawa, M. (2001) 'Inferring web communities through relaxed cocitation and dense bipartite graphs', *Proceedings of 2001 Data Engineering Workshop (DEWS 2001)*, pp.301–310.

Richard, B. and Dragos, C. (2009) *A Combinatorial Approach to Matrix Theory and its Applications*, Chapman & Hall/CRC Press, New York.

Richard, C. and Kamini, N.M. (2002) 'Evolutionary and historical aspects of the burden of malaria', *Clin. Microb. Rev.*, October, Vol. 15, No. 4, pp.564–594.

Richards, W.D. and Seary, A.J. (2003) *MultiNet*, Ver. 4.38 for Windows, Simon Fraser, Burnaby.

Saad, Y. (2011) 'Numeric methods for large eigenvalue problems', *Society for Industrial and Applied Mathematics*, 2nd ed.

Satya, K., Peter, G., David, R. and Martha, S. (2007) 'Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review', *PLoS Pathog.*, Vol. 3, No. 10, pp.1361–1371.

Shiow, C.T. (2010) 'Malaysia to release GM mosquitoes into the wild', *Science and Development Network*, 2 November [online] http://www.scidev.net/en/news/malaysia-to-release-gm-mosquitoes-into-the-wild.html (accessed 20 November 2010).

Timothy, F., Rita, C., Joan, R., Stephen, M., David, R. and Terry, Y. (2009) 'Using satellite images of environmental changes to predict infectious disease outbreaks', *Emerging Infect. Diseases*, Vol. 15, No. 9, pp.1341–1346.

Tom, B. and Gerardo, C. (2009) 'The reproductive number Rt in structured and non-structured populations', *Math. Bio. Sc & Eng. J.*, Vol. 6, No. 2, pp.239–259.

WHO (2009) 'Do all mosquitoes transmit malaria?', *WHO Online Q&A*, 11 December, Geneva [online] http://www.who.int/features/qa/10/en/index.html (accessed 11 December 2011).

WinWeb (2011) *WinWeb Crawler 2.0 Installation Site*, 4 April [online] http://www.winwebcrawler.com/index.htm (accessed 4 April 2011).

Xu, G., Yanchun, Z. and Lin, L. (2010) *Web Mining and Social Networking: Techniques and Applications*, Springer, New York.

Ying, D., Erjia, Y., Arthur, F. and James, C. (2009) 'PageRank for ranking authors in co-citation networks', *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 11, pp.2229–2243.

Yu, P.S., Han, J. and Faloutsos, C. (Eds.) (2010) *Link Mining: Models, Algorithms, and Applications*, Springer, New York.

Zhang, X., Li, Y. and Liang, W. (2010) 'C&C: an effective algorithm for extracting web community cores', in Yoshikawa, M. et al. (Eds.): *DASFAA 2010*, pp.316–326, LNCS 6193.

Zhao, B., Qian, W. and Zhou, A. (2010) 'Towards bipartite graph data management', *2nd International Workshop on Cloud Data Management (CloudDB2010)*, 30 October, pp.55–62, Toronto, Canada.

Zhou, Y., Lu, L. and Li, M. (2012) 'Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity', *New Journal of Physics*, Vol. 14, No. 033033, p.17.