on how to evaluate the quality and trustworthiness of online resources[6-11]. Pew Internet and American Life Project's Report [12] found that about a third of the Pew respondents felt the need to check the accuracy and reliability of the information they read. To the best of our knowledge, our proposal; on the use of the WKTM to evaluate the trustworthiness of web sites is different from other approaches.

II. AINI'S KNOWLEDGE BASES

Another significant difference between this research and other research on CAs is the domain knowledge model. Dahlbäck and Jönsson [13] stressed that the domain model represents the structure of the knowledge which comprises a subset of general knowledge. Such systems normally are comprised with two subcategories: the *traditional/narrow domain* or *domain-specific*, and the *open-domain*. In the traditional domain, systems attempt conversational fluency based on limited domains of expertise. ELIZA [14], for example, simulates a Rogerian psychotherapist, and its implementation is commonly known as DOCTOR and PARRY[15]. DOCTOR and PARRY's domain was restricted to paranoid hospital patient expressions. SHRDLU [16] is another program simulating a CA which is able to interact within a simple world knowledge of "blocks". SHRDLU was an entry in an early Loebner Prize competition, where the evaluation was based on the restricted tasks [17]. However, in the Fifth Annual Loebner Prize Contest in 1995, the Loebner prize criteria were changed to include unrestricted domains [18], requiring computer entries to converse indefinitely with no topic restrictions.

Hence, it is understood that general purpose CAs are not necessarily able to answer questions on a specific domain subject. On the other hand, domain-specific systems lack the flexibility to handle common sense questions. To overcome the above limitations, we proposed the Domain Knowledge Matrix Model (DKMM) [19] as shown in Figure 1. The data server layer serves as storage for data and knowledge required by the system. This is where AINI's conversational knowledge bases are stored. It is well understood that true intelligent action requires large quantities of knowledge. Such a reservoir of knowledge can be harvested from the internet and deployed in the domain matrix knowledge bases' architecture. This forms the basis for the construction of large-scale knowledge bases to be used as the engine for intelligent conversation systems. AINI is the mechanism used to manage the knowledge and to provide appropriate answers to the user.

AINI's DKMM incorporates several knowledge subjects. This is analogous to the consultation of expertise knowledge from multiple experts. For example, a *sales* knowledge domain should contain expertise on how to improve sales. However a sales person is expected to have a wide range of common sense which enable CAs have ability to engage the potential customer in general conversation. Hence, an intelligent system should also incorporate open-domain knowledge to handle general or generic questions. By including multiple domain knowledge bases within

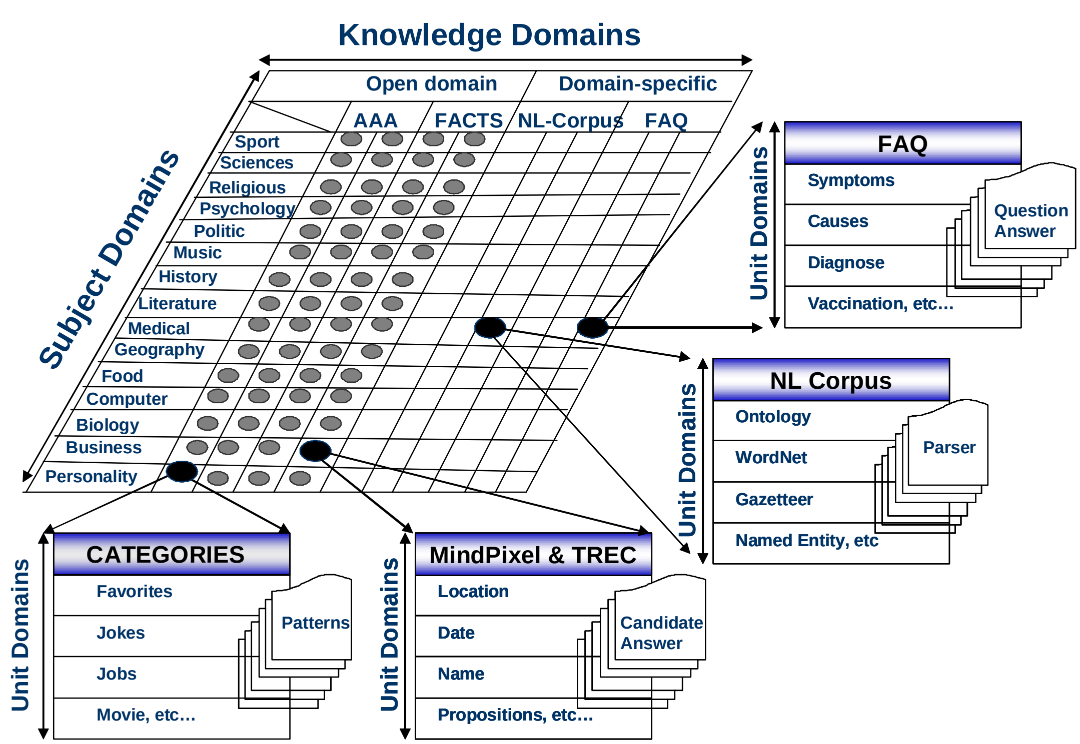


Figure 1: Domain Knowledge Matrix Model (DKMM)

AINI's single knowledge domain, the proposed AINI will be able to hold "meaningful" and prolonged the conversations with the users.

In this proposed DKMM [19], both the open-domain and domain-specific knowledge bases are predefined in the agent's knowledge. These modules are used to support the various knowledge levels at the agent brain tier. Depending on the user's input, the agent will respond or switch from one level to another in the agent brain. While the system is capable to communicate with the user beyond the knowledge domain, there are cases where the system will exhaust its capability to answer the queries. In such case, the system will attempt to divert the focus back to the current topic of interest by responding with some predefined random statements. The purpose is to direct the user's attention back to the system's domain-specific state. Hence, AINI will attempt to "cycle" between the six levels of information processing within the agent brain tier supported by the various knowledge modules in the agent knowledge tier.

A way to view the proposed DKMM is given in Fig. 1. In this approach, the knowledge base of the AINI can be considered as a collection of specific conversation domain units. Each unit handles a specific body of knowledge used during the conversation between AINI and the user. The knowledge can be seen as arranged in the vertical columns making up the open-domain or domain-specific knowledge. In addition, specific subjects are shown in the horizontal rows. For example, in the open-domain knowledge, the subject units will cover topics such as personality, business, biology, computers, etc. In this research, our focus is on the subject of medicine; and in particular, the bird flu pandemic. Therefore, additional bird flu domain knowledge is being incorporated in the domain-specific row "medical", and column NL-Corpus extracted using Web Knowledge Trust Model (WKTM).

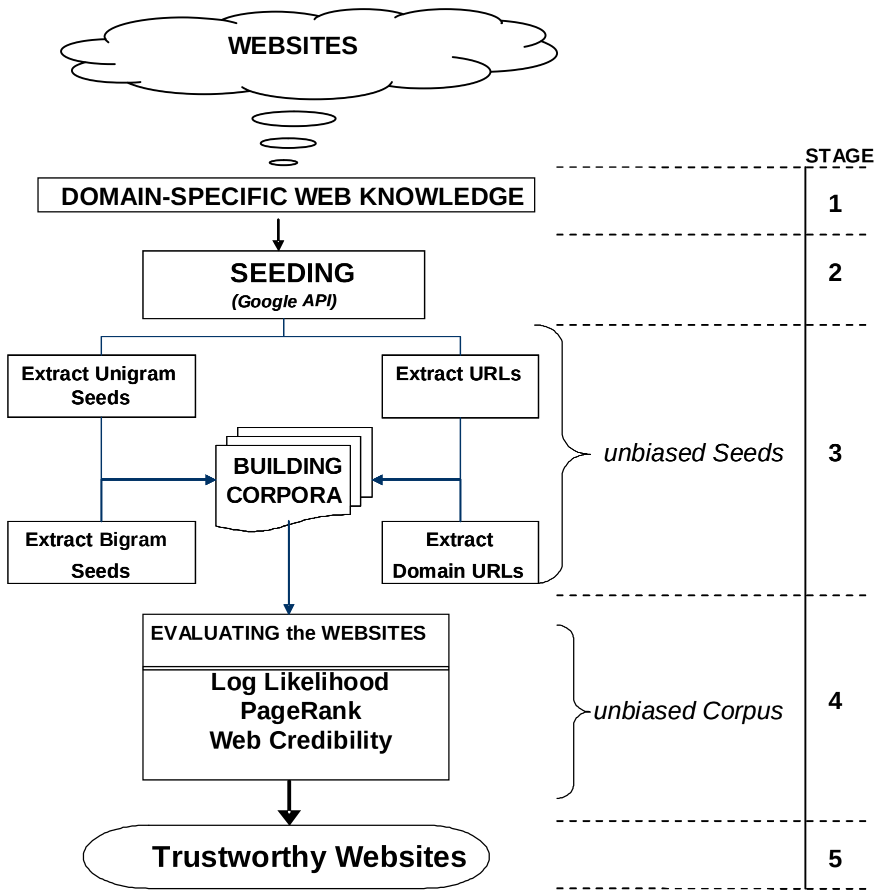


Figure 2: Web Knowledge Trust Model (WKTM)

III. WEB KNOWLEDGE TRUST MODEL (WKTM)

The objective of the Web Knowledge Trust Model is to provide solutions that will empower developers to adhere to the procedure described in Figure 2. It is expected that the model is also applicable to other application domains. The procedure outlined below is set out to address the question of *"how to select the most trustworthy domain knowledge from existing online web documents?"* The WKTM procedure can be divided into five stages. First, the target of the web domain knowledge to be extracted is determined. For this study, pandemic Bird Flu is the focus of the domain knowledge. In the second stage, a number of seeds are used in an iterative algorithm to bootstrap the corpora using unigram terms from the web. The process then proceeds to extract bigram terms based on the final corpus and unigram terms extracted in the previous phase. Once the sets of domain URLs have been collected, they are then submitted as queries to the search engine via Google API (Application Program Interface)[1].

All the downloaded URLs will be used to build a final domain corpus. In the fourth stage, the corpus obtained are evaluated using Log Likelihood, Google's PageRank algorithm[20] and Stanford's Web Credibility criteria [8]. Finally, the top five most trustworthy websites will be selected and extracted by AKEA.

IV. WEB DOMAIN KNOWLEDGE

In this experiment, the Bird Flu pandemic is the focus of the domain knowledge base. In current times, pandemic flu is becoming increasingly important in the research for real-world applications. The Head of philanthropy at Google, Larry Brilliant, has also described his vision on how information technology can be used to fight pandemics [21]. However, as the Web becomes increasingly chaotic and has strong possibility of misleading and inaccurate health information, the Web could become harmful to the unwary users. Selection of trustworthy web pages is therefore becoming an important factor in ensuring the long-term viability of the Web as a useful global information repository. The detailed descriptions of the subsequent stages in the WKTM are given in the following sections.

V. SEEDING

The purpose of this stage is to select the corpus as a data acquisition resource for building the CA's knowledge bases. Our aim is to create a "balanced" corpus of Web pages which contains relevant key words and documents of a given domain. For the purpose of seeding, we use words from the general training corpus, British National Corpus, (BNC)[2]. The BNC corpus consists of a 100 million word collection of samples of written and spoken language from a wide range of sources. It is designed to represent a wide cross-section of British English from the later part of the 20th century in both spoken and written forms. Since this research focuses on the Bird Flu pandemic, the initial seeds should come from its generic term derived from "bird" and "flu". From these seeds, we made a query to the online "specialized terminology" lists from the health information website MedLinePlus[3] Medical Dictionary. We found "bird flu" is related to "avian influenza". With these four seeds, we sent a query to the BNC online corpus and we obtained "virus" as an additional seed. From the bigrams observation, the seed "virus" occurred 19 times in "flu virus" and 11 times in "influenza virus". Finally, we collected the five terms "bird", "flu", "avian", "influenza" and "virus" for use as initial seeds for our investigation.

Table 1. Comparing hit results from BNC and Google's Corpora using the set of seeds

SEEDS	BNC		Google	
Unigram	**Freq of BNC Counts**	**%**	**Freq of Web counts in '000s**	**%**
bird	3869	63.14%	14,400	33.13%
flu	573	9.35%	4,790	11.02%
avian	45	0.73%	1,360	3.13%
influenza	145	2.37%	2,120	4.88%
virus	1496	24.41%	20,800	47.85%
Bigram				
bird flu	1	3.23%	602	46.45%
avian influenza	0	0.00%	180	13.89%
flu virus	19	61.29%	206	15.90%
influenza virus	11	35.48%	308	23.77%

Once the seeds have been obtained, a comparison is made between the BNC corpus and Google's large-scale corpus from public Web pages. The purpose of the comparison is to determine whether the BNC corpus is covering similar terms or updated information as in the web. A comparison of the results from the two sources is shown in Table 1.

In Table 1, the Freq of count is the number of returns from searching BNC corpus and Google. As expected, the counts are much larger from Google than from the BNC. The frequency of the total web counts from Google is 7,093 times larger than the BNC counts in the case of the unigrams. As for the bigrams, the Google Web counts are 41,806 times larger. These data were collected on 12th December, 2007. This evaluation demonstrates that BNC is not small in terms of the frequency counts due to a smaller corpus as compared to Google. In addition, it can also be observed that the distribution of the seeds in the unigrams and bigrams are not similar. For instance, "avian influenza" as a scientific term for "bird flu" is not included in the BNC; whereas in the Google corpus, this term accounts for 13.89% of the returns from the seed queries. In addition, the colloquial term "bird flu" only occurred at a frequency of 3.23% in the BNC whereas in the Google corpus, the same term occupied almost 50% of the returns. From this exercise, it can be assumed that Google takes into account of the continual increase in the page volumes and scale-up its corpus accordingly. On the other hand, BNC has not been able to keep up with newer terms such as "avian influenza" as indicated in Table 1. This also proves that BNC is insufficient by itself to provide the most updated information on any domain as in this case. However, as an initial stage in establishing the seeds for further query, the BNC has its merit as a training corpus. On the other hand, the Google

[1] http://www.google.com/apis
[2] http://www.natcorp.ox.ac.uk/
[3] http://www.nlm.nih.gov/medlineplus/mplusdictionary.html

retuned over 600 thousands of web counts in the case of the seed word “bird flu”. This again makes any attempt to extract the knowledge from all these pages impossible. This therefore leads to the need to establish a more refined corpus and in particular, to acquire knowledge from trustworthy sites. The process is described in the following section.

VI. BUILDING THE CORPUS

In this stage, a domain-specific corpus on pandemic Bird Flu is built using crawling approach. According to Broder et al. [22], crawling typically starts from a set of “seeds”. In this case, the seeds are obtained from the previous stage and consist of the five terms “bird”, “flu”, “avian”, “influenza” and “virus”. The crawling process consists of (a) fetch a page, (b) parse the page to extract all linked URLs, (c) for all the URLs not fetched previously, repeat (a)–(c).

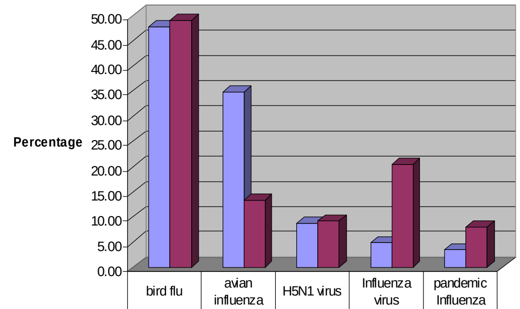
Normally, the crawling action will stop at some maximum value as limited by the Google API. For free service, Google limits the maximum number of queries to 1,000 per user per day. In this research, the number has been set as 10 URLs per search.

The Google API is used to analyze the result rankings for several queries of different categories using statistical tools in the BootCAT Toolkit [23]. The corpora are essential resources for knowledge professionals who routinely work with specialized domain knowledge. BootCAT toolkit implements an iterative procedure to bootstrap specialized corpora and terms from the web requiring only a list of “seeds” as input. Bootstrapping typically starts from a set of seeds randomly combined, and each combination is used as a Google query string. The top ‘*n*’ pages returned for each query are retrieved and formatted as text. These are the seeds which are expected to represent the domain under investigation. We make a first query to the Google search engine via Google API to extract the first corpus, and then extract new seeds from this corpus to build the second corpus [22].

Several important search parameters have to be controlled by the user, such as the number of queries to be issued for each iteration, the number of seeds combined to build a query, and the number of pages to be retrieved for each query, and so forth. The first step of this phase is to extract a list of single- and two-word connectors from the corpus (unigrams and bigrams). During this phase, we found an additional seed called “H5N1”, which was frequently connected with other seeds in the corpus. Hence, we added “H5N1” as the sixth seed to the seed set.

The second step is to retrieve the final URLs to build the final corpus. For simplicity and to avoid bias, only HTML and English pages are included. For each of the six seeds, BootCAT sends a query to obtain the number of URLs related to the seeds. The number of the final URLs returned is 1500 pages. After discarding the duplicated and broken URLs, the URL’s related to the domain under investigation is 1428.

A link analysis is applied to these sites under each domain name. If two domain names are linked with inbound and outbound connections, they are considered to be in a neighborhood. Only the domains which are included in the neighborhood are then selected. A few pages from each domain are then randomly chosen and concatenated into a document. After post-crawl cleaning, a corpus of 2,641,660 tokens is determined. This becomes the “Pandemic Corpus” in this research.



	bird flu	avian influenza	H5N1 virus	Influenza virus	pandemic Influenza
Pandemic Corpus (%)	47.82	34.90	8.86	4.95	3.47
Google Large-scale Corpus (%)	49.01	13.31	9.19	20.44	8.05

Seeds

Figure 3: Comparing distribution of seed words between the smaller set data Pandemic Corpus with the Google Large-scale Corpus

In order to verify the usability of this smaller corpus, it needs to compare the distribution of returns with respect to the larger Google corpus. This is shown in Figure 3. Although this corpus was created using a smaller set of sample seeds, it has a similar distribution as Google as seen from the figure. Hence it proves that the unbiased method as described in this proposal yields a similar coverage as Google. This leads to the next stage of evaluating the selected corpus and towards establishing trusted and reliable domain knowledge bases.

VI. EVALUATING THE PANDEMIC CORPUS

Before one attempts to carry out an evaluation, it is necessary define the term ‘trustworthiness’ associated with websites based on the credibility reports by [8] and [24]. Trustworthiness, a key element in the credibility calculus, is defined by the terms ‘reliable’, ‘truthful’, ‘unbiased’, and so on. Authority, another dimension of trustworthiness, is defined by terms such as ‘authorized’, ‘reputable’, ‘accredited’, ‘credentialed’ and ‘empowered’. The word “authority” often indicates a government or an educational institution controlling the contents of a site. The authority dimension of trustworthiness associates with reputable organizations. Combining these two dimensions, this suggests that highly trustworthy websites will be perceived to have high levels of credibility [8, 24] and authority. Based on these premises, this research is aimed at selecting the specific elements of a website that would lead to its consideration as a ‘trustworthy’ website. The elements proposed are based on Log likelihood ratio, PageRank and Web Credibility. They are described as follows.

A. Log Likelihood Ratio

In order to verify that the smaller pandemic corpus extracted by the proposed model is compatible to the large Google Corpus, the Log likelihood ratio is used as a quantitative assessment. The likelihood-ratio (LL-ratio) approach is a statistical method in which a ratio is used to illustrate the coverage probability and accuracy within the confidence interval for two corpora. The higher LL-ratio value indicates similar coverage probability even with small sample sizes [25] [26] [27].

The bigrams-based version of the log likelihood measure in the *Ngram* Statistical Package (NSP)[4] is used. In Table 2, the high LL-score values indicate the most important similarities between the two corpora for the coverage of the seed words. The results show that the proposed approach produces a confidence interval for the seed words with a nearly exact coverage probability and a high level of accuracy for the small pandemic corpus as compared to the Google large-scale corpus.

Table 2. Log-likelihood Ratios for Pandemic Corpus vs Google large-scale Corpus

Bigram	Pandemic Corpus	Google Large-scale Corpus in'000s	LL- Score
bird flu	12640	27,100	+106266.72
avian influenza	9223	7,360	+95698.31
H5N1 virus	2342	5,080	+19635.16
Influenza virus	1307	11,300,000	+ 7387.20
pandemic Influenza	918	4,450	+ 6233.06
Total Corpus	2,641,660	1,024,908,267	

B. PageRank

Evaluating a website manually is not an easy task. Another approach to use the Google's PageRank algorithm [20]. PageRank is a unique democratic process relies on the nature of the Web by using the web's vast link structure as an indicator of an individual page's value. It is the core algorithm of the Google's search engine. The algorithm is a complex and automated method which makes human tampering with the PageRank results extremely difficult. It should be noted that Google does not sell placements within the results thereby maintaining the democratic and unbiased nature of the search results. In this research, PageRank is used as one of the criteria to evaluate the trustworthiness of the websites based on link analysis. A similar application of link analysis is the evaluation of the quality of an academic work by analyzing the amount of citations. The number of backlinks to a given page gives some approximation of a page's importance or quality. PageRank extends this idea by not considering the links from all pages as equal. The algorithm also normalizes the final value to a range of 0 to 10. PageRank is defined as following algorithm:

$$p_i = (1-d) + d\sum_{j=1}^{n}(I_{ij}/c_j)\,p_j \qquad (1)$$

Suppose we have n webpages. Let $I_{ij} = 1$ if page j points to page i, and zero otherwise let c_j equal the number of pages pointed to by page j (number of outlinks). The Google PageRanks p_i are defined by the recursive relationship where the parameter d is a damping factor which can be set between 0 and 1. ie usually set d to 0.85.

In this study, the selection of trustworthy websites starts with selecting of the initial six seed words: ***bird, flu, avian, influenza, pandemic*** and *H5N1*. Based on the 1,428 URLs returned from stage 3, a query is sent to Google's PageRank directory to determine their rankings. Figure 4 shows the results of the top 10 sites based on the PageRank scale. The least important site is one with a PageRank of 1. The most referenced and supposedly important sites are those with a P_i of between 7 and 10.

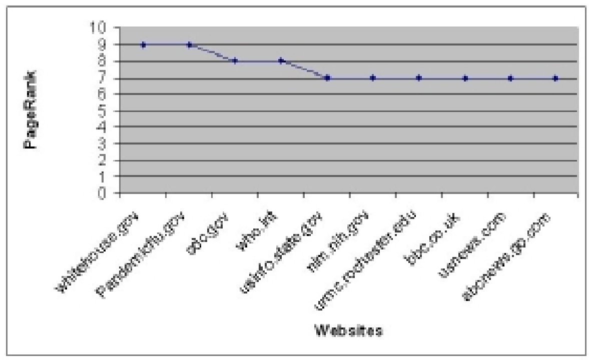


Figure 4: Pagerank values of Top 10 sites in the Bird Flu Domain

C. Web Credibility

This section presents the credibility of the top 10 websites related to this study assessed by a form of qualitative approach. After the PageRank results has been collected from the top 10 sites, a site is assigned with scores manually by experts based on the Web Credibility ranking criteria [8]. In this experiment, ten experts from the American Association of Webmasters in the web design field were asked to assess the credibility of these sites based on their professional judgement. The 'Top 10' sites collected from Google PageRank were then ranked according to their mean scores, highest to lowest. This ranking gives a general idea about which sites in this study have been found to be the most or the least credible by the users. When a more credible site was listed on the page, the site's score was given a point and the less credible site lost a point. Over the course of the study, each site was evaluated many times, gaining and losing points along the way. At the end of the study, each site received a final score, which was the average (mean) of all the scores it had received from the experts. The average value is the total number of points divided by the total number of times the site was ranked. If a site has a score of +1.0, it means

[4] NSP Package can be downloaded at http://search.cpan.org/~tpederse/Text-NSP-1.03/

the site is deemed to be credible by all participants. If the score is 0.0, it means the site was considered to be credible half of the time. Combining the three methods described, Figure 4 shows the results of the trustworthiness analysis for the top 10 sites related to the domain knowledge in this study.

VII. TRUSTWORTHINESS OF WEBSITES

The final set of URLs was further culled to include only selected sites attributed to regulated authorities. They are mainly government bodies, international organizations or educational institutions. All these organizations control and provide the contents of their respective sites. Once the seed set is determined, each URL's page is further examined and rated as either reliable or reputable. As shown in Figure 5, the selection is reviewed, rated and tested for connectivity with the trusted seed pages. The expert participants in the web credibility assessment exercise preferred websites that contain a great deal of information, instead of publicity news from the media such as BBC News, ABC News and USNews. These results also showed that the content or information factors were more important than design features in describing trusted or well-liked sites. In the current study, the final five websites cluster at the top of the web trustworthiness rankings are: pandemicflu.gov, whitehouse.gov, who.int, cdc.gov and nlm.nih.gov. All these highly credible sites were selected based on PageRank and credibility scale scores. These five top sites are clearly viewed by the expert participants as more credible than the other five sites in this study.

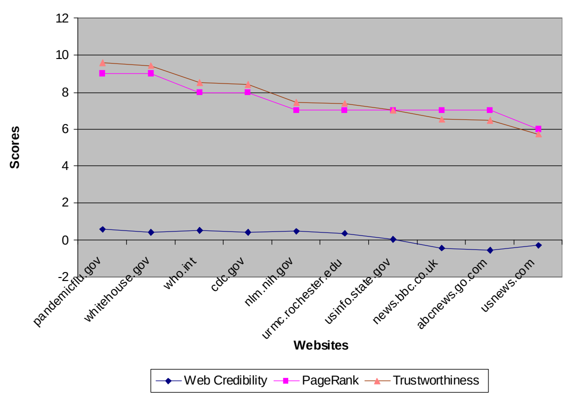


Figure 5: Comparing Trustworthiness of Top 10 Websites related to the Bird Flu Domain

The results support the proposal that the trustworthiness of websites is not only based on the PageRank and Web Credibility, but also the 'authority' of the websites which is not taken into account within the PageRank and Stanford Web Credibility criteria. There are other important factors in determining the 'reliable authority' of a site. They could be based on the site's history and the number of back-links to government agencies, education institutions, and international organizations. The more established and relevantly linked a site has, the more likely it could be considered as 'stronger' or 'more reliable'. This may effectively suggest the linked site has 'authority', 'reputability', 'empowerment' and 'credentials'. This work will be examined in future study. Finally, the top five URLs are then used as the main source of knowledge for AKEA to extract the pandemic related contents to build AINI's domain-specific knowledge base.

Currently[5] AINI's open-domain and domain-specific knowledge bases has more than 160,000 entries in the common sense stimulus-response categories. Of these, 100,000 came from MindPixel, 997 factoid questions from the TREC training corpus and 45,318 categories from the AAA knowledge bases. On the domain-specific knowledge base, AINI has about 10,000 stimulus-response categories extracted from trusted online documents based on WKTM. This makes up over 160,000 stimulus-response items in total. AINI also has 158 FAQ pairs of questions and answers, which have been updated using AKEA. In addition, AINI has also collected more than 52,890 utterances in conversations with online users since the first prototype of AINI was put online in the February 2006[19]. These utterances will be integrated into AINI's knowledge bases through supervised learning by domain experts. At present, AINI has learnt about 5,000 categories from conversations with online users. All of this combined knowledge has made up the total of 161,473 stimulus response categories in AINI's knowledge bases. To compare AINI with other systems, the original conversation programs such as ELIZA, written by Professor Joseph Weizenbaum of MIT, has only 200 stimulus response categories. ALICE Silver Edition was ranked the "most human" computer, and has about 120,000 categories, which include 80,000 taken from MindPixel as summarised in Table 3.

Table 3: AINI's Stimulus-response Categories

Domain Knowledge	Sources	Categories	%
Domain-Specific	NL Corpus	10,000	6.19
	FAQ	158	0.10
Open-Domain	MindPixel	100,000	61.93
	TREC Corpus	997	0.62
	AAA	45,318	28.07
Supervised Learning	Conversation Logs	5,000	3.10
	TOTAL	**161,473**	**100**

[5] Till 1 August 2007, AINI's have 161,473 stimulus-response categories in their knowledge base.

VIII. CONCLUSION

Based on the proposal and experiment described in this paper, the contributions of this research are:

1. The procedure of selecting trustworthy websites for building a conversation agent's knowledge bases is proposed.
2. A scheme for selecting a "unbiased seed set" for building a corpus has been presented.
3. A Web Knowledge Trust Model (WKTM) for determining reputable, credible, reliable and accountable websites is proposed.
4. Results of an evaluation based on 1,428 Bird Flu Pandemic websites crawled by Google API are presented and discussed. Some interesting statistics on the hit frequency, a significant data collection based on PageRank and Stanford Web Credibility are observed. The corpus is also used to evaluate the proposed WKTM.

These contributions indicate that this novel approach contributes towards the building of restricted CAs domain knowledge based on WKTM. The proposed model demonstrates the credibility of the web sites could be defined and is probably closer to a realistic expectation of trustworthiness. The URLs traces and data sets from this research are available on the Internet for future research. Another data collection phase is also planned in order to examine the application of the results presented here with a new set of domain knowledge. The future study will assess the robustness and comprehensiveness of the knowledge extracted from the web in addition to the trustworthiness issues.

Acknowledgements

This research project was funded by Murdoch University Research Excellence Grant Scheme (REGS), 2006/07

REFERENCES

[1] A. Franz and T. Brants, "All our N-gram are Belong to You," Google Machine Translation Team, 20 October 2006, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

[2] A. Kilgarriff and G. Grefenstette., "Introduction to the special issue on the web as corpus," *Computational Linguistics,* vol. 29, pp. 333–347, 2003.

[3] O. S. Goh, C. C. Fung, K. W. Wong, and A. Depickere, "Embodied Conversational Agents for H5N1 Pandemic Crisis," *Journal of Advanced Computational Intelligence and Intelligent Informatics,* vol. 11(2), pp. 282-288, 2007.

[4] O. S. Goh, A. Depickere, C. C. Fung, and K. W. Wong, "Domain Matrix Knowledge Model for Embodied Conversation Agents," in *5th International Conference on Research, Innovation & Vision for the Future (RIVF'07)*, Hanoi, Vietnam, 2007.

[5] O. S. Goh and C. C. Fung, "Automated Knowledge Extraction from Internet for a Crisis Communication Portal," in *First International Conference on Natural Computation* Changsha, China: Lecture Notes in Computer Science (LNCS), 2005, pp. 1226-1235.

[6] MadTek, "Evaluating Health Related Websites," MadTek, 19 april 2006, http://www.madtek.com/content/index.pl?iid=2837&isa=Category.

[7] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with TrustRank," in *Thirtieth International Conference on Very Large Data Bases*, Toronto, Canada, 2004.

[8] B. J. Fogg, L. Marable, J. Stanford, and E. R. Tauber, "How Do People Evaluate a Web Site's Credibility? Results from a Large Study," Consumer WebWatch & Stanford University, Yonkers, N.Y 2002.

[9] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen, "What Makes Web Sites Credible? A Report on a Large Quantitative Study," in *ACM CHI 2001 Conference on Human Factors in Computing Systems*, Seattle, WA, USA., 2001, pp. 61-68.

[10] R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," in *CHI 2006*, Montréal, Québec, Canada., 2006.

[11] WebWatch, "A Matter of Trust: What Users Want From Web Sites," Consumer Reports WebWatch, 1 May 2007, http://www.consumerwebwatch.org/pdfs/a-matter-of-trust.pdf.

[12] S. Fox and L. Rainie, "Vital Decisions: How Internet users decide what information to trust when they or their loved ones are sick," Pew Internet and American Life Project, Washington D.C. 2002.

[13] N. Dahlbäck and A. Jönsson, "Integrating domain specific focusing in dialogue models," in *Eurospeech'97*, Rhodes, Greece, 1997, pp.,2215-2218.

[14] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM,* vol. 9, pp. 36-45, 1966.

[15] K. M. Colby, *Artificial Paranoia: A Computer Simulation of Paranoid Processes*: Pergamon Press Inc., 1975.

[16] T. Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," *Cognitive Psychology,* vol. 3, 1972.

[17] S. M. Shieber, "Lessons from a Restricted Turing Test," *Communications of the ACM,* vol. 37, pp. 70-78, 1994.

[18] H. G. Loebner, "In Response [to Shieber 1994a]," *Communications of the ACM,* vol. 37, pp. 79–82, 1994.

[19] O. S. Goh, C. C. Fung, and A. Depickere, "Domain Knowledge Query Conversational Bot in Instant Messaging (IM)," *Knowledge-Based Systems,* vol. 21, 2008.
[20] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems,* vol. 30, pp. 107–117, 1998.
[21] M. L. Baker, " Open-Source Initiative Targets Bird Flu," Ziff Davis Publishing Holdings Inc, 29 April 2007, http://www.eweek.com/article2/0,1895,1963157,00.asp.
[22] A. Z. Broder, M. Najork, and J. L. Wiener, "Efficient URL Caching for World Wide Web Crawling," in *WWW 2003*, Budapest, Hungary., 2003.
[23] M. Baroni and S. Bernardini, "BootCaT: Bootstrapping corpora and terms from the web," in *Fourth Language Resources and Evaluation Conference*, 2004.
[24] B. J. Fogg and H. Tseng, "The Elements of Computer Credibility," in *CHI99 Conference on Human Factors and Computing Systems*, 1999, pp. 80-87.
[25] S. S.Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics,* vol. 9, pp. 60-62, March 1994 1994.
[26] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics,* vol. 19, pp. 61-74, 1993.
[27] S. L. Piao, G. Sun, P. Rayson, and Q. Yuan, "Automatic extraction of Chinese multiword expressions with a statistical tool," in *Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, 2006, pp. 17-24.

(Revised online publication: 21 June 2008)