ICED13
INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED13
19-22 AUGUST 2013, SUNGKYUNKWAN UNIVERSITY, SEOUL, KOREA

USING WEB CRAWLER TECHNOLOGY TO SUPPORT DESIGN-RELATED WEB INFORMATION COLLECTION IN IDEA GENERATION

Zhihua WANG, Peter R N CHILDS, Pingfei JIANG
Imperial College London, United Kingdom

ABSTRACT
Effective information gathering in problem and task related fields with which designers or design teams may not be familiar is a key part of the design process. Designers usually consult with subject experts to access expert information. An Effects Database system that includes design-related effects to provide ready access to expertise at any stage within the design process can be used to prompt areas to consider and explore. To maintain the efficiency of the system, its data must be regularly updated and new effects populated from the open source knowledge base. Web crawler technology has integrated into an information gathering and analysis system to rapidly mine design-related information from published data sources in order to update an effects database for use in design. This paper describes the effectiveness and efficiency of the system for updating the database. Comparing with manual information collection, the test results demonstrate that this system can dramatically increase the efficiency on selecting design-related information from un-restricted internet sources.

Keywords: design methods, information management, database, web crawler

Contact:
Zhihua Wang
Imperial College London
Mechanical Engineering
London
SW7 2AZ
United Kingdom
z.wang09@imperial.ac.uk

1 INTRODUCTION

In practical design activities and opportunities, designers often face a design task that relates to fields they are relatively unfamiliar with and the knowledge accumulation they have is not specific to the design task (e.g. see Lidwell et al., 2003). As the accumulation of fundamental knowledge in a field is important for the promotion of new insights in that field, in order to compensate for shortages of knowledge accumulation in problem-related fields, designers may choose to acquire the relevant knowledge for themselves, with the associated time commitment, or consult experts for guidance and thereby benefit from available prior experience (e.g. see Amabile, 1996; Csikszentmihalyi, 1996).

Wang and Childs (2013) carried a survey to investigate reasons for ineffective information gathering in unfamiliar problem related fields. The survey revealed two relevant indications: when designers face a design problem related to unfamiliar knowledge fields, they have to spend a significant proportion of time to define the correct knowledge scope for the problem, which causes time pressure; when designers consult experts for help in knowledge gathering, experts may be stymied by ineffective dialogue with limited understanding of subject subtleties or design context. In normal daily life, people typically have a limited amount of attention to learn knowledge in diverse fields, the quantity and complexity of which have increased with time (Csikszentmihalyi, 1996). When designers face a problem that lies in unfamiliar fields, they may spend a long time in obtaining a basic accumulation of knowledge. Therefore, consulting experts for guidance is often an expedient solution. In order to improve access to expert information, a database of design-related effects, named the Effects Database has been implemented to enhance the information scope in idea generation (Wang and Childs, 2013).

However, as the complexity and variety of modern knowledge fields continually increases with the development of knowledge, new effects or principles may be proposed and some existing ones may be refined. The main purpose of the Effects Database is to provide problem related effects and principles to assist designers to rapidly define the knowledge scope of design tasks. To maintain the effectiveness of the system, the data in the system must be regularly updated and new proposed effects identified from published data sources and included.

The aim of this paper is to describe an information gathering and analysis system integrating a web crawler technique that periodically refreshes the data in the Effects Database and evaluate its performance. In section 2, the Effects Database System is briefly introduced. After illustrating the web crawler technique, several open source web crawlers are compared in section 3. Section 4 provides a description of the work flow of information gathering and analysis system. The effectiveness and efficiency of the system is explored in section 5.

2 EFFECTS DATABASE

The Effects Database system includes effects and principles that are collected from existing domains, including physics, chemistry, geometry, mathematics and psychology. As creativity can be described as "the ability to invent or develop something new of value" (e.g. see Childs, 2004), for a design, the "new value" can be of societal or financial benefit or indeed both of these. The majority of designs stem from new incorporations of earlier inventions and an invention in one field can be transformed as the integration of coordinated fundamental technical effects in that field and other associated fields (Henderson, 1991; Orloff, 2006). The effects provided by the Effects Database System aim to assist designers undertaking improvements for existing designs, such as integrating new functions, or improving the performance of components and sub-systems as well in the formation of new concepts to deliver specific functions. The system as implemented by Wang and Childs (2013) has 128 physical effects, 78 chemical effects, 28 geometric effects, 47 psychological principles and 46 design principles. For each effect or principle, a definition, book or article reference and a web reference have been developed and selected. An example is shown in Table 1.

The main function of the Effects Database is to provide problem related effects and principles to assist designers to rapidly define the knowledge scope of design tasks. The investigations of Howard (2010) revealed that many companies ran idea generation sessions to directly provide product ideas. The system can be integrated into the standard idea generation process as used in companies and industries and the working process of the system is shown in Figure 1.

Table 1. An example effect.

#	Physical principle	Definition	Book or Journal reference	Web reference
3	Thermal – electrical phenomena	The direct conversion of temperature differences to electric voltage and vice versa.	Serway, R., Jewett, J. W. Jr., 2002. *Principles Of Physics: A Calculus – Based Text.* 3rd ed. Press: Thomson Learning.	http://www.metacafe.com/watch/1944815/you_gotta_see_this/

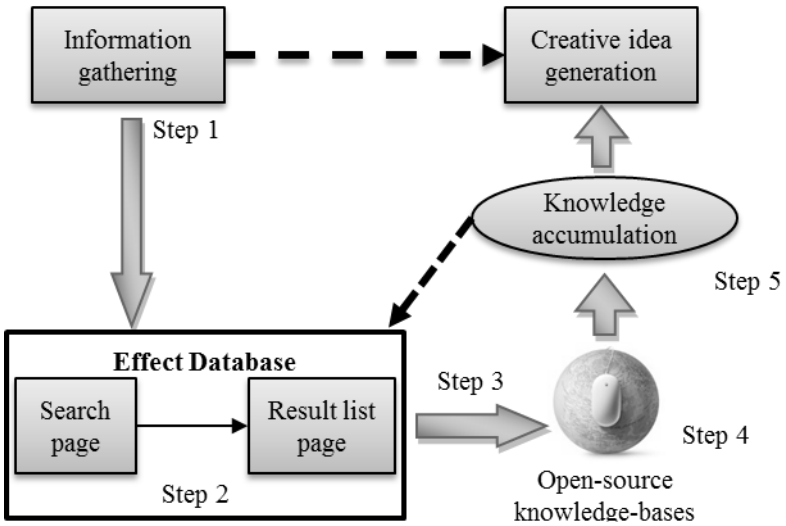

Figure 1. Using the Effects Database in the early stage of the idea generation process

The approach developed is as follows.

1. *Concluding problem related keywords (Step 1):* Based on the information and facts from previous stages in the design process, some keywords are proposed. A keyword should be a simple word that is related with the design problem.
2. *Searching the database* (*Step 2):* In this step, once designers enter a keyword into the database, the background programs search keyword related effects or principles in the database and then list results obtained in the results list page.
3. *Exploring information for each effect identified (Step 3):* This step is for designers to briefly examine the information of each effect or principle in the results list page including its definition, book or article reference and web reference. After finishing examination of some or all of the keyword related results, designers can revisit Step 2 to enter another keyword.
4. *Exploring more information for the specific effects identified (Step 4):* During step 3, designers can refer to the book reference and the web reference of an effect or principle to explore more information from open-source knowledge-bases.
5. *Enhancing the knowledge accumulation in problem related fields (Step 5):* After all the keywords have been entered and keyword related results examined, designers have an indication of the scope of knowledge they may need to explore in the problem related fields. Moreover designers could "inform or suggest to" the experts that these effects or principles were related with their design tasks and ask for related specific knowledge guidance even though the experts may not understand the design tasks. Thus, the problem related effects and principles could be selected to enhance their knowledge accumulation.

Sometimes, the knowledge guidance from the effects database or experts may stimulate designers to conclude new keywords or replace previous keywords by more appropriate ones. In this situation, designers may repeat the database searching process using the new keywords. The use of the database has been illustrated by application to a series of design tasks, indicating its suitability for promoting expert relevant suggestions.

3 WEB CRAWLER TECHNOLOGY

Use of search engines such as Google, Yahoo, Bing or Baidu to explore information from the internet (Prieto et al., 2012) is commonplace. The central part of the search engine is web crawler technology. A web crawler is a programme that can browse the World Wide Web in an orderly sequence. During

the crawling process, it creates a copy of all the webpages visited for later processing. Castillo (2004) concluded a typical structure of standard web crawlers as shown in Figure 2.

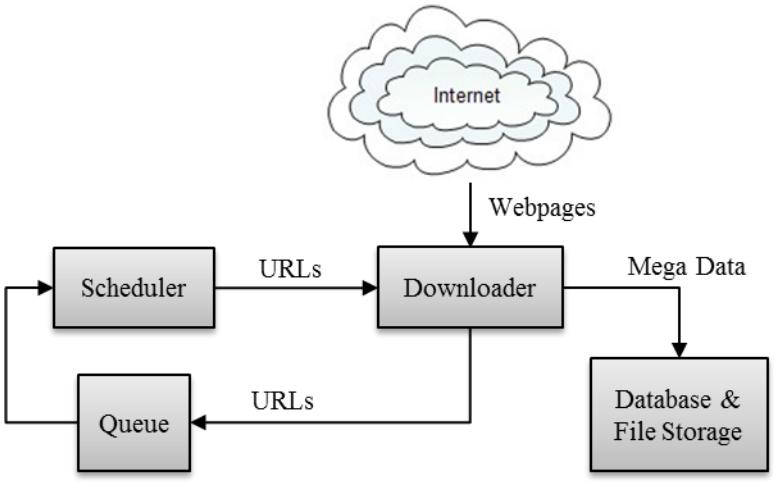

Figure 2. The high level architecture of a standard web crawler (see Castillo 2004)

The crawler starts with a list of URLs to visit. As it visits these URLs, it identifies all hyperlinks in accessed webpages and adds them to the list of URLs for future visit (called the Queue). URLs from the queue are recursively visited according to a set of policies, which are as follows:

- *Selection policy*: states which pages to download
- *Re-visited policy*: states when to check for changes to the pages
- *Politeness policy*: states how to avoid overloading web sites
- *Parallelization policy*: states how to coordinate distributed web crawlers

Many commercial search engine companies have produced their own web crawlers (examples shown in Table 2) to provide up-to-date data. Because these search engines are commercial, the algorithms and architectures of these crawlers are kept confidential and locked to prevent others from reproducing the work. Thus, these commercial crawlers are unable to be used by public without permission and licensing.

Table 2. Some published crawler architectures for general purpose crawlers

Search Engine	Yahoo!	Bing	Google	Fast Search & Transfer ASA
Crawlers	Yahoo! Slurp	Bingbot	Googlebot	FAST Crawler

In addition to proprietary crawlers, e.g. Table 2, there are many general open-source crawlers available. Many of these are licensed under the GNU General Public License (GPL), which means they are free to be studied or modified. Su et al. (2005) investigated the features of some open-source crawlers. The results are summarised in Table 3.

Table 3. Some characteristic features of a selection of general open-source crawlers under the GPL

Crawlers	Features
ASPseek	An internet search engine, written in C++ using the STL library.
GNU Wget	A command-line-operated crawler written in C. Used to mirror Web and FTP sites.
Heritrix	Written in Java and specifically designed for web archiving.
HTTrack	Written in C. Creates a mirror of a web site for off-line viewing.
Nutch	Coded in Java. Data is written in language-independent formats.

After exploring the features of these open-source crawlers, HTTrack was selected as a basis to update the Effects Database system. The merits and advanced features identified are included in the following list.

1. Websites can be downloaded from the Internet to a local directory. Recursive building of all directories, getting HTML, images and other files from the server to a local computer.
2. It retains the relative link-structure of original site. When users open a page of the mirrored website in the local webpage browser (e.g. Internet Explorer, Chrome and Safari) to view the site from link to link, they may 'feel' a similar experience as viewing the webpage online.

3. It is fully configurable and provides an integrated user guide system, which is a useful tutorial feature for novice users.

4 INTEGRATING HTTRACK INTO THE INFORMATION GATHERING AND ANALYSIS SYSTEM

The high level flow chart of the information gathering and analysis system is shown in Figure 3. This includes two main parts: Webpage Downloader and Webpage Content Analyser.

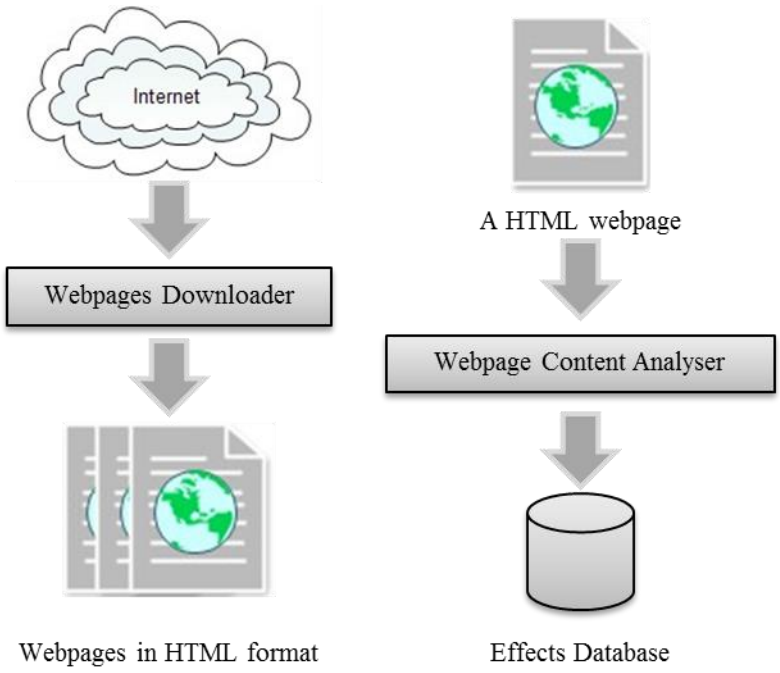

Figure 3. The high level architecture of the information gathering and analysing system

The Webpage Downloader downloads webpages from internet and converts the format of these webpages into Hypertext Markup Language (HTML). Subsequently, the Webpage Content Analyser analyses the main content of each downloaded webpage and writes selected information into the database of the Effects Database. The HTTrack is integrated into the Webpage Downloader.

4.1 Webpage Downloader

HTTrack has been implemented two versions: a GUI version and command-line-operated version. For the GUI version, users can define the setups through windows operation frames and dialogues. However, due to the various setups for websites with specific format structures, the user has to manually set corresponding setups for different websites, such as URLs, results save path and downloading rules. If there is a large number of candidate websites, the setup work may be time consuming. Furthermore, the setup work for a website has to be repeated when the same website is downloaded again. Fortunately, this open source programme also provides a command-line-operated version, which has been integrated with the Webpage Downloader. Website related setups can be supplied through a Command Prompt. Therefore, the corresponding setups for each website could be set and saved in a configure file and the setup work for all websites does not need to be repeated.

The flow chart of the Webpage Downloader is shown in Figure 4. Initially, the programme reads the candidate websites list from database. This list includes the name of the websites, accessed URLs, downloaded webpages saving directory and other specific setups of each website. Then, for one website, the HTTrack caller will scan the running processes of the computer to check whether the HTTrack programme is free or not. Although HTTrack can be multi-process running, to simplify further maintaining work, here it is used as a single-process programme. If the last HTTrack is still running, the programme will recursively check its status every 1000 ms until it is released. The programme will start a new downloading process and set the corresponding setups for the downloading process of this website through the Command Prompt. The webpages on this website will be accessed and downloaded by HTTrack through the URLs provided. After this step, the

programme will move to the next candidate website in the websites list. If there is another website, the programme will go back to the HTTRack caller and start a new loop.

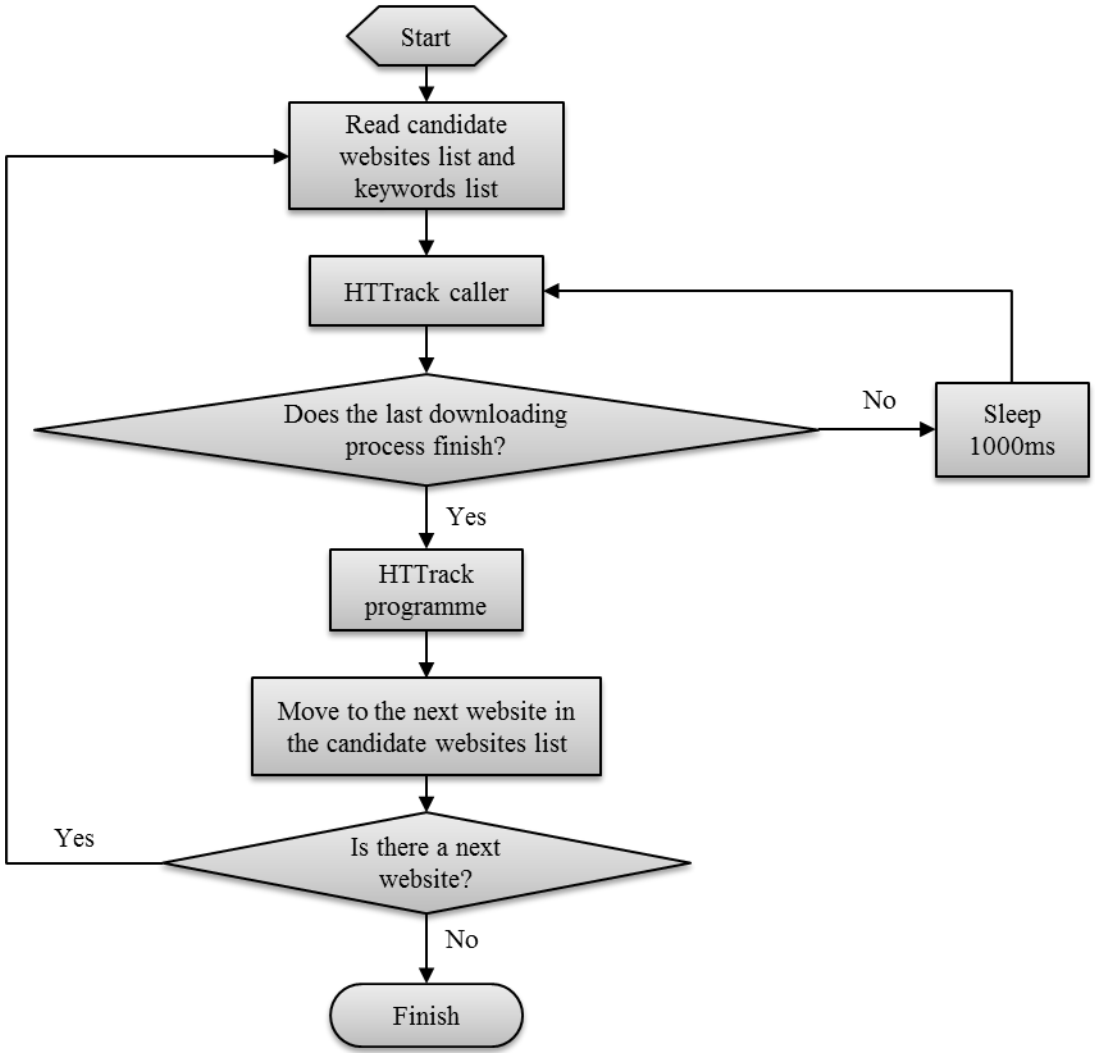

Figure 4. The flow chart of Webpage Downloader

4.2 Webpage content analyser

In the Effects Database, each effect is provided with several related keywords. One example is shown in Table 4. The effects "Thermal conduction" is provided with 11 related keywords in this example.

Table 4. An example effect with related keywords

Effects	Keywords
Thermal conduction	Energy, transferred, heat, atomic, scale, energetic, molecules, continuum, energy, colliding, molecules

After finding all the effects related with the effect under consideration in the database, there are 213 generic keywords in total in this particular case (shown in Table 5).

Table 5. Generic keywords concluded from the data in the Effects Database

Generic keywords
vibration, electromagnetic, wavelength, ferromagnetic, currie, magnetization, paramagnetic, reluctance, magnetize, atomic, molecules, continuum, colliding, radiation, melting, boiling, isothermal, condenser, microwaves, dielectric, collision, electrostatic, superconductors, photons, transformation, insertion, specular, reflection, diffuse, scattered, electrons, absorption, resultant, nucleus, hadron, lepton, gamma, quantum, scateriing, fusion, fission, vector, magnetism, radioactivity, nucleons, pascal, oscillations, inertia, polarized, electrostriction, perpendicular, capillary, solvent, membrane, porous, electromotive, elevation, spinning, entanglements, polymer, sparking,

modification, longitudinal, vapour, bubbles, gradient, dispersed, amplitude, susceptible, adsorption, entropy, magnitude, friction, creep, vessels, pulses, kinetic, bremsstrahlung, pendulum, gravity, propagate, electromagnetic, proportionality, electromotive, superconductors, resistance, planar, incompressible, compressibility, cohesion, coefficients, phenomenological, tensor, electron, vacuum, superimposed, bloch, adhesion, plastic, refractive, pockels, kerr, photomagnetic, beams, polarizing, inductor, reactance, capacitor, atoms, amplification, resonance, recoil, momentum, transverse, proton, demagnetization, foams, propagation, oscillation, ferroelectric, dipole, plasma, phosphorescent, lattices, acoustic, elasticity, birefringence, cooper, quasi, spins, phonon, components, splitting, polarization, isotropic, carbon, benzene, refringent, semiconductor, endothermic, exothermic, chemiluminescence, luminescence, enthalpy, oxidant, oxygen, synthesis, combustion, metallic, polymeric, isolation, oxidation, aerosol, droplets, hydrogen, ionic, covalent, interstitial, alkali, deoxidization, anion, adsorbents, condense, monomer, helium, electrolytes, hydrogen, alloys, emulsion, bromide, vapourisation, solidification, molecules, clustered, covalent, synthetic, amorphous, heterogeneous, homogeneous, lattice, repel, electroplating, chromium, alloy, acid, mordant, ions, radicals, polymer, pigments, indicator, intermolecular, entity, radioactive, organism, oxides, macromolecules, inhomogeneous, cylindrical, diffraction, inhomogeneities, restoration, reactant, intermediate, anodic, cathodic, compliant, convex, crank, propulsion, paraboloid, hyperbolic, saddle

The Webpage Content Analyser works on webpages individually. It will examine all the downloaded webpages and select ones related with the generic keywords in Table 5. The corresponding flow chart is shown in Figure 5.

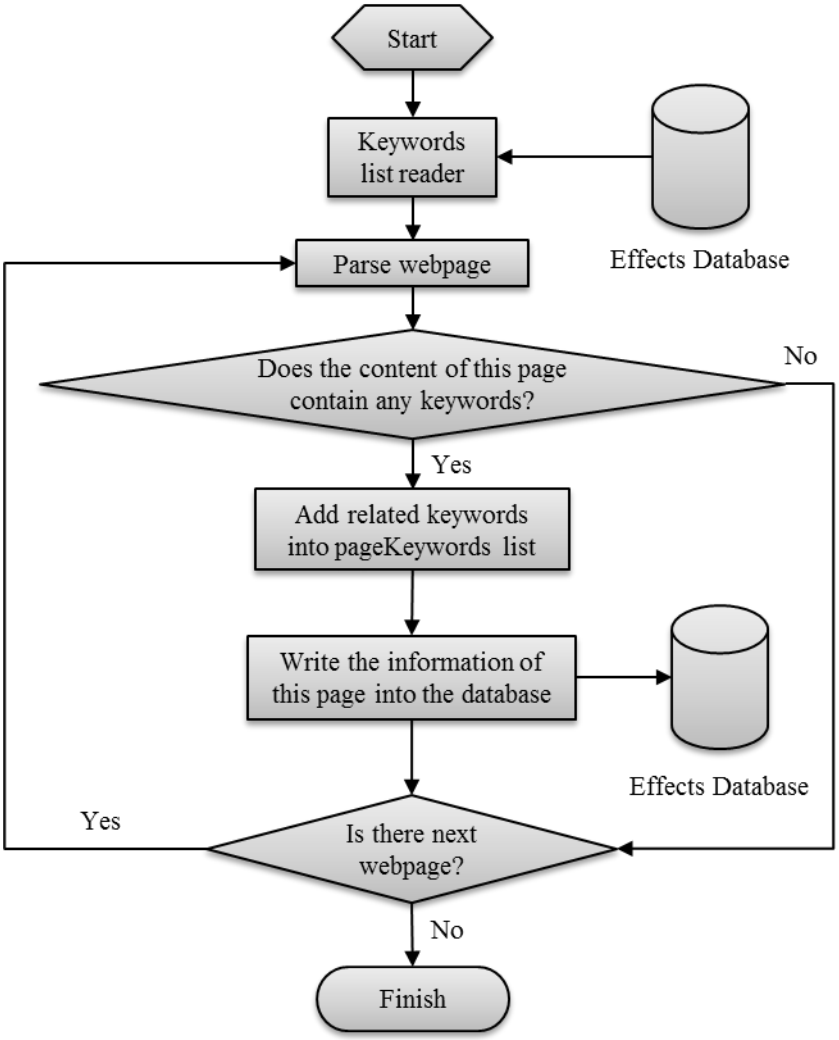

Figure 5. The flow chart of the Webpage Content Analyser

Firstly, the programme reads the keywords list from the database. This list contains all the generic keywords. Then, the source code of the page will be parsed and its main content extracted. For the

next step, the content of this webpage will be matched with each keyword in the list. If the content contains any keywords, this page will be saved and add all related keywords into the pageKeywords list of this webpage. Following this, the page information and related keywords in the pageKeywords list will be written into the database for further processing. If no keyword is contained in the keywords list of a webpage, this webpage will be ignored. After this, the programme will move to the next webpage and parse it until all webpages are analysed. In this way, all of the downloaded webpages will be analysed and keywords-related ones will be selected and stored into the database.

5 SYSTEM PERFORMANCE ANALYSIS

5.1 Performance test

The main function of the Webpage Downloader and Webpage Content Analyser is downloading webpages from candidate websites, extracting valuable information and writing them into the database of the Effects Database. Based on the investigation from Henderson (1991), effects should be collected from the conclusions from existing domains or from design knowledge from existing designs. Therefore, the candidate websites should have high relationships with technology breakthroughs. Additionally, those websites should be open accessed by the Webpage Downloader.

The survey from the 11th annual Higher Education – Business and Community Interaction reviewed that UK universities contribute over £3.3 billion value to the economy and society through applying knowledge into business (HEFCE, 2012). This may indicate that universities in the UK, as in other nations, are pioneers in new knowledge research. Usually, the news release websites of universities will publish the latest research breakthroughs online. Therefore, these websites could be employed as candidate websites for system performance analysis. The news release websites of 5 well-known universities in the UK (shown in Table 6) were chosen as the candidate websites. For each website, only the news released in the year 2012 was downloaded.

Table 6. Candidate websites for the system performance analysis

Name	URLs
University of Cambridge	http://news.admin.cam.ac.uk/news/2012
University of Southampton	http://www.southampton.ac.uk/mediacentre/news/2012/
University of Bath	http://www.bath.ac.uk/news/2012
University of Manchester	http://www.manchester.ac.uk/aboutus/news/archive/
University of Loughborough	http://www.lboro.ac.uk/service/publicity/news-releases/2012

Table 7. Detail of the downloading process

Name	Time cost (s)	Pages downloaded	Size (MB)	Average rate (kB/s)
University of Cambridge	158	197	1.67	10.80
University of Southampton	85	207	2.21	26.57
University of Bath	179	256	4.06	23.79
University of Manchester	227	383	5.88	26.52
University of Loughborough	108	243	2.87	27.87
Total	977	1286	16.69	

Table 8. The selected results of the Webpage Content Analyser

	University of Cambridge	University of Southampton	University of Bath	University of Manchester	University of Loughborough	Results	Cost time (s)
Raw webpages	197	207	256	383	243	1286	18.9
Selected webpages	26	66	43	96	47	278	

The setups including the URLs of these 5 websites, the year, 2012 in this casem and other limits were edited and saved in the configure file of the system. Then, the Webpage Downloader programme downloaded webpages from the 5 websites sequentially according to their setups. All downloaded webpages were saved in different subfolders but in the same parent folder. The internet environment is

10 Mbps The detail information of the downloading process is recorded and shown in Table 7. In the following step, the Webpage Content Analyser started to analyse these webpages individually and select out keywords-related ones. The selected results are given in Table 8.

5.2 Result Analysis

From Table 7, the time cost on downloading 1286 webpages is 977 s. The time cost on webpage content analysis process is around 19 s. Thus, the total time cost on the entire two processes that selecting 278 keywords-related webpages from the five websites is 1000 s (<17 minutes). Weinreich et al. (2008) revealed people normally spend 70 seconds on reading a webpage with around 1000 words content. The total time cost on online reading all downloaded webpages would be 90020 s, which is around 1500 minutes. Regardless of including the time for judging whether the webpages are keywords-related or not, this time is nearly 90 times than the time used by the system.

After manually analysing the 278 webpages, 88 webpages of them were confirmed as scientific related and could be used to update some data in the Effects Database. For example, the University of Cambridge had launched several low carbon techniques to protect living environments. The effect concluded from these techniques would be valuable for designers to deal with design tasks related with low carbon requirements. The University of Southampton investigated phase change materials during rapid heating. This found could update the effects related with thermal transmission which would benefit designers from solving design tasks related with heating. The University of Bath reported on a new bamboo-based construction material, in detail. This material could update the effects related with compound materials or multi-layer materials. The University of Manchester had achieved breakthrough research on graphene. Its new features could be added into the database as effects related with electrical device designs. And the University of Loughborough pioneered new 3D printing techniques.

Apart from the confirmed 88 webpages, the system also indicated the remaining 190 webpages were keyword-related. After exploring the content of these pages, the reasons identified were as follows. Some webpages introduced the biographies of professors, whose research areas were keyword-related. These webpages might be valuable for designers to find specific experts for design-related knowledge guidance. Some webpages published prizes awarded to students or researchers, whose achievements are keyword-related. Designers might benefit from this portion of information for ideas generation stimulated by these achievements. There were also some announcements of scientific activities which were keyword-related. The remaining 190 webpages contained less valuable information when compared with the confirmed 88 webpages. Excluding those keywords-related 278 webpages from the total downloaded webpages, some of the remaining webpages may also contain valuable information. Considering the amount of time spent on manual webpage analysis, it was decided that it would be inefficient to manually search these webpages.

Therefore, based on the test data and the above analysis, the information gathering and analysis system could dramatically increase the efficiency on selecting keywords-related webpages from information sources on the internet. Some of the principles used in undertaking the manual selection process can be incorporated within an algorithm to further reduce the time or need for a manual selection process altogether and this represents an area of on-going work.

6 CONCLUSION

In this paper, an information gathering and analysis system including two parts, a Webpage Downloader and Webpage Content Analyser, is proposed to update the data of an Effects Database system which has been produced to aid designers in generating creative ideas in short periods of time compatible with modern business. With scientific development, some data in the Effects Database may need to be provided with new definitions or include new effects and data to maintain its effectiveness and efficiency in providing problem related effects and principles to assist designers rapidly define the knowledge scope of design tasks. A web crawler technique was integrated with the Webpage Downloader to mine webpages from candidate websites. The Webpage Content Analyser was used to analyse the content of those downloaded webpages and select keywords-related ones to update data or provide new data entries for the Effects Database. The performance of the system was tested by providing five candidate websites which were the news release websites of five well-known UK universities. 1286 webpages were downloaded and 278 keyword-related webpages were selected by the software. The webpages were manually explored and 88 of them were confirmed scientific related

and could be used to refresh the data of the Effects Database. Comparing the time cost on selecting keywords-related webpages by a manual approach and that of the software system, the system had much higher efficiency on selecting keyword-related webpages from internet.
Based on the data, the final confirmed webpages was 7% of the downloaded webpages (88 out of 1286). If the number of candidate websites increases to 50 or 100, such as the websites of the world top 100 universities and research institutes, the number of keywords-related webpages will considerably increase. In the test, only the news in 2012 of the five websites was downloaded in the test. If more years were included, the number of selected webpages will dramatically increase.
The system could be set to run periodically. In this way, the data in the Effects Database can be regularly updated to contain contemporary scientific research breakthroughs as well as established effects and principles.

REFERENCES

Amabile, T.M. (1996) *Creativity in context*, Westview Press.
Castillo, C. (2004) *EffectiveWeb Crawling*, Ph.D. in Computer Science, University of Chile.
Childs, P.R.N. (2004) *Mechanical Design*, 2nd ed., Elsevier Butterworth-Heinemann Ltd., Oxford.
Csikszentmihalyi, M. (1996) *Creativity: flow and the psychology of discovery and invention*, Harper Collins Publishers.
HEFCE (2012) *Higher Education – Business and Community Interaction Survey 2010-2011*
Henderson, K. (1991) Flexible Sketches and Inflexible Data Bases: Visual Communication, Conscription Devices, and Boundary Objects in Design Engineering. *Science, Technology, & Human Values*, Vol. 16, No. 4, pp. 448-473.
Howard, T.J., Culley, S., and Dekoninck, E.A. (2010) Reuse of ideas and concepts for creative stimuli in engineering design. *Journal of Engineering Design*, Vol. 22, No. 8, pp. 565-581.
Lidwell, W., Holden, K., and Butler, J. (2003) *Universal principles of design*, Rockport Publisher.
Orloff, M.A. (2006) *Inventive Thinking through TRIZ: A Practical Guide*, 2nd ed., Springer Berlin Heidelberg.
Prieto, V.M., Alvarez, M., Lopez-Garcıa, R., and Cacheda, F. (2012) A Scale for Crawler Effectiveness on the Client-Side Hidden Web. *Computer Science and Information Systems*, Vol. 9, No. 2, pp. 561-583.
Su, C., Gao, Y., Yang, J., and Luo, B. (2005) An efficient adaptive focused crawler based on ontology learning. *Hybrid Intelligent Systems*, Rio de Janeiro, 6-9 Nov, HIS '05, pp. 73-78.
Wang, Z. and Childs, P.R.N. (2013) Using design-relevant effects and principles to enhance information scope in idea generation. *Research into Design*, Chennai, 7-9 Jan, Springer, pp. 137-149.
Weinreich, H., Obendorf, H., Herder, E., and Mayer, M. (2008) Not quite the average: An empirical study of Web use. *ACM Trans*, Vol. 2, No. 1, pp. 1-31.