A Decentralised Approach to Information Retrieval for a developing country like Bangladesh

Harnessing Education and Technology to End Poverty

Hammad Ali, Nafid Haque

BRAC University, Bangladesh.

e-mail: {hammad2099, nafid99}@yahoo.com

Abstract – In this paper, we talk about a decentralised information retrieval system which would be suitable for the developing countries that face the problem of limited bandwidth. In this paper we came up with an implementation that uses the existing technology in a novel manner to meet the specific needs of users in such countries. We considered the infrastructural limitations of a developing country like Bangladesh and thus the solution presented here performs just as well in similar situations anywhere else in the world. Our work has been tested for the Bangla language and the same procedure can be applied easily for any other language in any part of the world where there is need for such a system. We had to pick and choose from a set of software packages for one that would best serve our needs. We also had to take into consideration user convenience, for which we had to keep in mind the diverse demographics of people that might have need of such a system. Finally, we came up with the system with all the desired features.

Introduction

Bangladesh, the country where our case study was conducted, is a developing country in South East Asia. Like most other developing countries, digital divide is also a serious problem here. While the big cites are enjoying broadband Internet technology, a large number of the rural areas are yet to be connected to the telephone networks let alone have access to Internet. Nevertheless, people in these areas are always in need for information. The type of information they need ranges from medical or legal knowledge base to the current market situation all over the country. A centralised knowledge base would be of no use because there is no convenient way to distribute this information uniformly throughout the country. In a very essential sense, access to information is a form of empowerment. In a developing country like Bangladesh, decentralisation is a precursor to uniform distribution of information. There is no way we can expect the rural people to produce their best unless we at least give them a level field as far as information is concerned. A situation where the city-dweller knows more than someone in the countryside can only lead to further increase in poverty. For poverty alleviation the first thing that needs to be made sure is that one class cannot manipulate another based on the disparity of information available to them.

Let us take a look into the type of information most needed in a country like Bangladesh. Firstly, there is a great need for robust medical database in the rural areas. This stems from the fact that the number of physicians in rural areas is mostly not sufficient to cater to actual needs. The physicians themselves often suffer from lack of resources, and this often results in malpractice. If we could have a trained info-mediary (information intermediary), equipped with a comprehensive medical database, it would not only help the physicians but also let the patients enjoy greater accountability. Another important document that would be of great help to people in the rural areas is the Penal Code of Bangladesh. People in these areas are often not very clear about their rights and what actions they might be able to demand from the Government or Law Enforcement Agencies in the country. Again, if the info-mediary is, to some extent, trained in legal issues, he might be able to use the Penal Code, maybe along with documentations of similar incidents in the past, to guide an individual in the right direction. Of course, this would mean that an info-mediary would have to be selected taking into consideration the kind of data he or she will be dealing with. A single individual cannot be adroit at both medical and legal issues, at least not enough to make a real difference. So we would have to ensure that the information

is managed by someone who is in a position to understand this data and put it to best use. Lastly, there is also need for information about the current market condition of the country. This information is particularly useful to farmers or fishermen who wish to sell their product to middle-men, who can then sell it in the big cities. Unless these people know the actual price of what they are selling, there is a very good chance that they will be ripped off by the middle-men who will try to maximize their own profit. This in turn will only further deteriorate the poverty situation. Thus, it is integral for the development of the country that these information be available readily to the masses.

Over the last decade or so, computers have become a very common tool in the area of storing information and accessing this information with convenience. They have become our preferred method of social interaction, work and most importantly, storage of data. Paper documents firstly have the disadvantage of a very short shelf-life, a definite bottleneck if we take into consideration the need for archiving newspapers and magazines for National Libraries and University Archives. Furthermore, the difficulty of extracting information from paper documents makes a further point in the argument for the efforts to convert these documents in digital form. The bigger the size of the data we have in digital form, the greater becomes the need for an efficient Information Retrieval system for this body of data. Without a better way to look up relevant data, the best we would be able to do is sequentially browse through all of this data. As can be easily seen, such a method would be inefficient to the extreme and would not serve our biggest purpose – convenient access to information regardless of the language in which this information is stored, and regardless of who is trying to get hold of this information.

The question of language highlights a very important point. Not so long ago, computer data was almost exclusively in English. In recent years, though, there has been a rise in the use of other languages with computers. Since the learning of the English language proved to be a bottleneck in the process of familiarizing people with computerized systems, there has been a greater frequency of using one's native language with computers. People are constantly trying to incorporate their native language into the use of computing in their daily endeavours. Bangla is one such language. Furthermore, Bangla is ranked between the fourth and the seventh most widely spoken language in the world and has nearly 200 million native speakers (Languages 2006)**.** However, this position of Bangla, and the extent to which it is used in written communication, is not reflected accurately if we only take into consideration the amount of Bangla corpus made available in digital form. Nevertheless, in recent times Bangla newspapers and magazines have come to terms with the power of sharing information and are gradually devoting more attention towards creating and maintaining their own websites. From this there is an automatic rise in the need for a Bangla search engine. This is particularly true for newspaper websites, which have extremely dynamic content but at the same time also need to ensure easy access to their archives. Apart from the Bangla documents available over the Internet there are documents in Bangla stored locally and used to get the information required at times of need. The most common example of a considerable amount of Bangla data residing in local machines would be the data repository that certain organizations want to make available to people in the rural areas of the country. What these entities intend to do is make a large body of data available in portable media, and then send these out to remote areas of the country where people might be in need of this information. Ensuring this equitable access to information is absolutely integral to the development of a country like Bangladesh. People in the rural areas of our country are often in need of information regarding specific issues and as already mentioned, all of these data can easily be made available in portable media that can be sent out to information kiosks in remote parts of the country. However, with this collection of information comes the need for efficiently searching through it based on a set of relevant keywords. Unless we could retrieve relevant information in the shortest amount of time possible, this large body of data would often be rendered useless or at best suboptimal. In fact, it would even mean that the larger the data collection, the more inefficient its use.

Towards finding a solution to this problem, what we are trying is not to coin a revolutionary new technology, something that the world has never seen before. We are more interested in *harnessing available technology for the sake of national development*. People are always concerned about ensuring human rights in a diverse number of sectors, ranging from education and medical facilities to

– in more recent years - the rights of practising one's own religion and philosophy in life. However, little has been done to ensure equitable access to information for people from all walks of life. Today, we are in the era of Information and Communications Technology. Just like before this there have been upsurges in the fields of industries, electricity or even the advent of computers as a tool, this is the era of the upsurge in Information and Communications Technology for Development. In today's world, the most influential men are those who have the most convenient access to the largest body of information. As stated in the WSIS Geneva 2003 Declaration, "our common desire and commitment to build a people-centered, inclusive and development-oriented Information Society, where everyone can create, access, utilize and share information and knowledge, enabling individuals, communities and peoples to achieve their full potential in promoting their sustainable development and improving their quality of life, premised on the purposes and principles of the Charter of the United Nations and respecting fully and upholding the Universal Declaration of Human Rights" (WSIS 2006).

Computers today have become a common household tool in developed countries. These nations enjoy the latest in modern technology, and it would not be an exaggeration to state that often, all the information in the world that they might need is just one click away for them. Now, it is a matter of common knowledge that this access to information is not equitable in all nations of the world. Furthermore, in most developing countries of the world, this infrastructure is not equally available even in all parts of the same country. Digital divide has become a matter of great concern in these countries. In some countries, the situation is such that in the cities people are enjoying the facilities provided by the latest and the best in communications infrastructure, whereas in other parts of the country people do not even have telephone facilities, let alone access to the Internet. For instance, in a developing country like Bangladesh, the Internet is a novel concept in most parts of the country. Ensuring Internet access all over the country is a faraway dream, something that might yet take decades. Even in the areas that do have Internet facilities, the quality of service varies widely within the span of only a few tens of kilometres. What we want is an alternative, one that can be achieved in a much shorter time but perform just as well. So *in this paper we propose an infrastructure for a decentralised approach towards information retrieval.* While our objective is to find an implementation that works with any language that is currently being used with computers, for the purpose of our case study we selected the Bangla language. *We believe that harnessing the available technology in a novel way can take a lead in alleviating poverty.*

Background Study

Extracting information from a collection of data is not a new concept. Many proprietary and open-source search engines are capable of this task. The search engines that support Unicode can easily be used for searching for a query string in any language supported by Unicode. One of the leading on-line search engines today is Google which has recently come up with an application called Google Desktop (Google). Google Desktop indexes the data stored on a local machine and allows an interface – identical to that of their web search engine - to search through the data with all the features of its on-line search engine. There are on-line search engines other than Google, like Yahoo (Yahoo) and MSN (MSN). However, since these leading search engines are all proprietary products of their respective developers, they cannot be customized to meet specific user needs. Apart from these proprietary search engines there are many open-source engines freely available that can be customized to meet specific user needs. Lucene (Lucene) is one of the most popular and mature open-source search engines available at present. Nutch (Nutch) is an environment built upon Lucene that gives the user all the features of a search engine. Vicaya (Vicaya) is a search engine built on top of Nutch and Lucene. Vicaya can reside by itself on any portable media, like a CD for instance, and enable searching through its contents without any prior installation on the machine. Our paper proposes a system similar to Vicaya but with additional features to make searching in Bangla or any other language easier for the user, regardless of their background. We have also studied different open-source content management systems to organize diverse text contents.

Methodology

Over the last decade or so, information has become one of the most coveted resources in the world. People from all over the world are in need of some form of information, and the Internet has become the grand avenue of sharing information. Towards this end of sharing information, a common language like English is a big help. People nowadays feel the need to learn efficient communication in English not just for the sake of learning the language, but because it has become the demand of their everyday work. However, there is another perspective to the issue. While English is absolutely necessary for communication on an international scale, relying on it too heavily can actually hamper the process of development within a country where English is not the native language, or even a very commonly used language. For instance, in a country like Bangladesh, the majority of the inhabitants are not very comfortable with the use of English. Trying to get them to learn sufficient English would be prohibitive in terms of the expense, time and human resources that would be needed. The better solution would be to provide these people with some way to access information in their own language, the one that they are comfortable with. From this spurs the need for making data available in Bangla, and then making access to this data as convenient as it is in the case of English.

The reason behind our focus on local machine as well as the web for our Information Retrieval System stemmed from the fact that there is a greater demand for searching Bangla text from a local machine or portable media than from the Internet. This is largely due to the lack of availability of Internet facilities in remote areas of the country. *In remote areas of the country people are in need of specific information. Keeping this in mind many Government and Non-government organizations have taken initiatives to setup information centers where a person can obtain their required information.* The organizations working with this concept wish to setup these information centers as close as possible to the people who need the information. However, there is no way at present to provide Internet facilities to these centers. Thus arises the need for providing all the information in some portable media like a CD/DVD ROM. Obviously, then comes the need to search through the corpus at regular intervals. At present, most of the available information retrieval/search facilities from a local media are based on the concept of raw string matching – probably the simplest and most inefficient method of searching. The problem with string matching is that it finds all and only the words that are spelled exactly like the query word. That is, it does not allow for the concept of fuzzy searches, words that are spelled very similarly, or words that sound phonetically similar and are very often confused for each other. Another limitation could be the fact that such searching techniques do not use the power of reverse indexing, something that has become very common in the arena of web search engines. This means that each time a query word is typed in, the entire corpus is searched all over again to find each and every instance of the word. As can be readily seen, this can be a great obstacle towards efficient retrieval of data.

So, we needed a state-of-the-art implementation that would perform optimally for data residing in the local machine. If the content changed frequently, we would also need to update the indexed database accordingly. This is important so that the new data can also be searched through, and relevant results shown to the user. If the data is essentially static, creating the index file need only be a one-time thing. We can collect all the data, index it and provide the indexed database in the same media along with the actual collection of data.

In Bangladesh, many private organizations have taken the initiative to setup small information centers in the remote areas of the country. One such initiative taken by Grameen Phone (GP), the largest mobile phone operator in Bangladesh, provides Internet access and other communications services to the rural people through its community information centers. D.Net (D.Net) is another such private organization working to give access to information for rural people. The objective of these projects is to ensure access to information to people in remote areas of the country, where Internet is not an option. For information retrieval from a local machine, we chose the D.Net projects Pallitathya (Pallitathya) and Abolombon (Abolombon). As mentioned above, there are several other such projects being managed by Grameen and other organizations. We chose D.Net since they are a representative of the standard mode of operation of such centres, and also because they allowed us access to some of

the content they had developed for the sake of this project. Details about these projects are given below:

Abolombon is a project geared towards creating awareness among rural people about governance and human rights issues. The project has several dimensions: the citizen's lack of awareness about their rights, lack of awareness related to the role and obligations of government institutions, lack of availability of information related to legal support and inadequate legal references for legal aid, among others (Abolombon). The project has developed substantive digital legal content in simple Bangla. This information has also been made available through a web portal (Abolombon). The legal contents are also available off-line through CDs distributed in project areas. D.Net delivers these contents through rural information centers in remote areas. A D.Net employee trained with the use of computers acts as the information intermediary and searches for precise, relevant data from these CDs for people who come to these centers looking for help and advice. This intermediary would be the end user of our system, the person typing in the query words, obtaining the results and presenting it in a useful form for the person in need of it.

Thus, to build a fully-functional Bangla information retrieval system, we chose Lucene and Nutch to be the base of our work (Nutch, Lucene). Our aim is to customize the open-source search engine Nutch so that it can search for Bangla text from portable media or the local machine. The implementation would also be able to do this regardless of the encoding method used. The system we have proposed has three different parts integrated into a common package. They are the content extractor, indexer and user-interface. The following sections will highlight each individual part of the package.

1. Content Extractor

A considerable amount of the documents available in Bangla are not encoded using the Unicode format. Since the Nutch search engine only supports Unicode, there is a need to convert any other encoding format to Unicode. We have used a simple Java program that crawls through the documents and tries to identify the encoding being used by analysing the Meta Tags, Font types etc. After identifying the type of encoding, the extractor program then converts only the text into its corresponding Unicode format, leaving out any graphics. Thus, given an URL, it would try to find out the encoding used for the site. If it is using Unicode then a crawl can be performed directly. However, if the data is encoded using some other encoding format, then it tries to find out more about the type of encoding being used through a set of heuristics based on stochastic modelling. Once the format is known, it can then use this knowledge to extract all the contents and convert them to its equivalent Unicode encoding format. Nutch can then simply crawl through this set of files and create the index. Searching can then proceed on the indexed Unicode database.

2. Indexer

Initially, we focused on the search engine API Lucene. Lucene is an open source project that provides the basic functionalities of a search engine without an interface (Lucene). It is written in Java and suitable for full-text search, especially in cross-platform scenarios. However, since Lucene is just an API it does not provide the framework for user input. So we moved onto Nutch, a package built upon Lucene. In addition to all the functionalities of Lucene, Nutch also provides an easy-to-use interface for user interaction (Nutch). We would also have the Unicode support that is crucial to make sure our implementation works with the language we picked. Of course, to exactly meet our needs we had to make certain changes to the default configurations for Nutch. We had to modify some of the parameters in the configuration files according to our needs and other issues such as the crawling depth. Once all this has been done, we can just run the crawl and create the database. Then this database could be used to search through the set of documents. We needed Tomcat (Tomcat, T_UTF8) server running on the machine to launch the user interface and enable searching through the database. With all of this ready, we could crawl any site, create the database and then search through the database. So we essentially had an information retrieval system ready for the Bangla language. If Unicode encoding was being used, we could just crawl through it and get the database ready. If some other encoding is being used, then we would need the extra step of using the content extractor to grab

the text content, convert it to Unicode and then crawl through that corpus. Nevertheless, the process of crawling and then searching through the database remains the same.

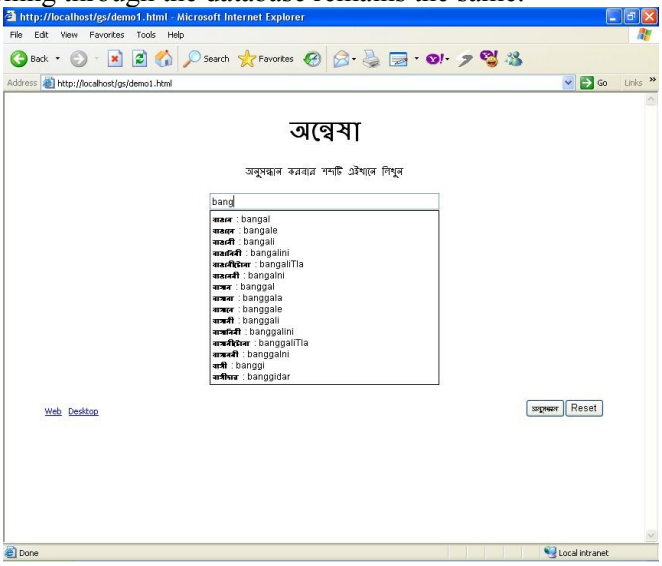

Fig. 1 Prototype Interface for our system

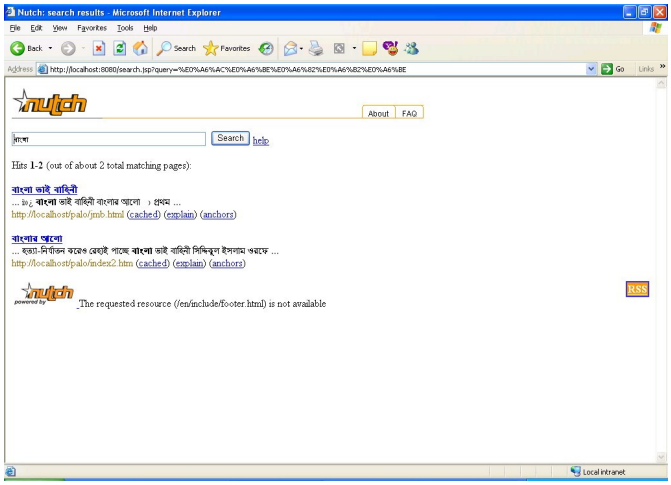

Fig. 2 Search results being displayed

3. User Interface

The section regarding user input needs a good consideration to make the entire project a success. A wide range of Bangla keyboard layouts are available and thus the users of Bangla keyboards are not focused on a single keyboard layout. Also, from recent trends it can be seen that people prefer to write Bangla phonetically using the English alphabets. Keeping all this in mind we designed our system so that it can be easily used by anyone regardless of the keyboard layout available to them. As the whole system is based on Unicode, users can use any Unicode supported Bangla keyboard to enter their search string. In case users do not have a Unicode supported Bangla keyboard, the system will provide the user an option to enter their search strings in a transliterated form (writing Bangla using English alphabets.) A JavaScript code has been developed and integrated with the system so that the user can enter their Bangla text in a transliterated form using English alphabets. The script code is also capable of suggesting the user with near matches of their input words, by consulting the index database for all the words that have this common prefix. This can be very useful in a situation where the user does not know the exact spelling of the word, a common scenario in the case of searching for names of individuals, among other possible entities.

With the three above stated segments for the proposed system, the entire system can be a very strong tool for information retrieval from a text collection of any size. All the contents need not be on the Internet but can reside on any portable storage medium or even the local machine (e.g. CD/DVD ROM). In case of searching data from a portable media, the contents need to be indexed first and then along with a live version of the system can be put on the storage medium preferred. Then, the storage medium (e.g. CD/DVD ROM) can be carried to any place and the data in it can be searched through just as one would in case of data over the Internet. If the contents of these documents may change periodically, then there is the need for re-indexing. In that case the system needs to be configured so that the data can be crawled at regular intervals and the index file updated accordingly.

Future Work

At present, the content extractor that we are using only works for text content. In future, we would like to hone this implementation further so that it can work for other, more diverse types of content. This is important, because most documents are likely to carry a lot of non-text information that could be just as important to the user. We want the user to be able to see all of this content while using our implementation and not just stay confined to seeing text. So we plan on working further with the content extractor in relation to our system. In addition, we want to enhance the system further, and go from an information retrieval system to an information extraction system. Such a system would have uses in the field of text summarization, text categorization using the presence of keywords or information extraction from an article or any information portal. Further, with a good amount of available data we might be able to build a strong knowledge base for the purpose of semantic analysis and content extraction. To improve performance of the system, we could do some more work on the cache policy that should be implemented for a system like this. We could also try and enhance the concept of grid sharing, such that each individual centre would serve as one node and a good number of these could somehow work together, sharing information in such a way so that each node does not need to store all information but only the subset most important to itself. Any other information could be obtained from the nearest centre with access to that particular information. In such a case, all of these nodes would come together to form a nationwide information backbone. Despite the infrastructural problems we have mentioned earlier, at least this degree of co-ordination is still possible between the centers. If we can utilize it to some degree, our solution might become at least a little more efficient, and thus enable us to disseminate information in a more effective manner.

Also, as a probable improvement for the user interface section, we could try to come up with an on-screen keyboard displaying the Bangla alphabets. The user would then simply have to click on the desired letter and it would show up on the screen. While typing in this manner would be slower, it eliminates the need for any sort of extra features at the user end, and provides a convenient input method that anyone would be able to use.

Conclusions

In this paper, we have tried to suggest a system that would make it possible to search through a large collection of data in any language within a feasible amount of time and expense, regardless of any particular features. For the purpose of our case study, we selected the Bangla language. We mostly paid attention to the question of information retrieval from portable media. We did not try to devise something completely original. *On the contrary, we were trying to suggest ways in which we could capitalize on available open-source technology to ensure elimination of digital divide in different parts of Bangladesh. We wanted to suggest an system that would make sure that access to important information within limited time becomes a right and not a privilege.* The main incentive behind our work was not the intention to come up with a system that none has thought of or worked on before. Rather, it was trying to contribute something to the mission planned by the United Nations and the International Telecommunications Union in their endeavour to ensure fair access to information for everyone regardless of their geographical location or other background information, as expressed in the Geneva Declaration of 2003 (WSIS 2003).

In a country like Bangladesh, the majority does not have access to the Internet. Furthermore, there does not seem to be any probability of ensuring Internet access in all parts of the country even within the next decade or so. So we tried to find a way to do the best we can *to harness the power of Information and Communications Technology for development*, even within the limitations that are imposed on us in the context of this country. We wanted a system that would ensure equal rights at least in the pursuit of information for people from all walk of life regardless of there locations. *Our work may not be ground breaking in terms of innovation or pioneering in a field, but we trust that it will prove to be very important for the sake of national development. We believe it was our responsibility to use our abilities, and available technology, to do something for the betterment of the nation*. Our work is still at an incipient stage, where we are trying out different approaches towards a solution to the problem of how technology play a role in development. In this paper, we report some of our findings and ideas so far. We sincerely hope that the system suggested here can go far towards this end. Whatever progress we make through working on this system for Bangla can then go a long way towards trying such implementations for other languages, and other countries that would be benefited by the use of modern technology as a means of poverty alleviation and equitable access to information. Advancement of technology is not an end in itself, and we have a responsibility to ensure that the people who benefit from technology are not only those who live in the big city and enjoy the latest and the best, but also those who are barely aware of the breakthroughs being made in the arena of Information and Communications Technology. They are in more need of the technological edge than anyone else. We hope the work we started can go a long way, to the day when people can truly enjoy equitable access to information and can witness the use of technology as a means for poverty alleviation.

Acknowledgements

We thank D.Net for giving access to some of their developed contents to test our system. We also thank Center for Research on Bangla Language Processing at BRAC University to give us the necessary guidance for conducting our work.

References

Languages (2006). The World's Most Widely Spoken Languages. Retrieved on October 2006 from http://www2.ignatius.edu/faculty/turner/languages.htm.

Google. Google Desktop, Retrieved on August 2006 from http://desktop.google.com/about.html.

Yahoo. Yahoo! Search, Retrieved on August 2006 from http://search.yahoo.com.

MSN. MSN Search, Retrieved on August 2006 from http://search.msn.com

Lucene. Erik Hatcher and Otis Gospodnetic, 'Lucene in Action', April 2006.

Nutch. The Official Nutch Website. Retrieved on August 2006 from http://www.lucene.apache.org/nutch.

Vicaya. Vicaya Website. Retrieved on August 2006 from http://vicaya.sourceforge.net.

D.Net. Development Research Network Website. Retrieved on August 2006 from www.dnet-bangladesh.org

GP. Grameen Phone Community Information Centers. Retrieved September 2006 from http://community.telecentre.org/en-tc/node/18324.

Pallitathya. A research program of D.Net on understanding information needs from a village perceptive. Retrieved on August 2006 from http://www.pallitathya.org/

Abolombon. A program of D.Net designed to improve access to legal information on governance and human rights issues. Retrieved on August 2006 from http://www.abolombon.org/.

N_Wiki. The Nutch wiki. Retrieved on August 2006 from http://wiki.apache.org/nutch/.

N_Tutorial. The Nutch tutorial for Version 0.7. Retrieved on August 2006 from http://www.lucene.apache.org/nutch/tutorial.html.

Tomcat. A step by step guideline on how to configure and use Tomcat, Retrieved on August 2006 from http://www.coreservlets.com/Apache-Tomcat-Tutorial/.

T_UTF8. Web log on enabling Tomcat to support UTF-8 Encoding. Retrieved on August 2006 from http://rollerweblogger.org/page/roller/20040415.

WSIS (2003). FAQ on the World Summit Information Society website. Retrieved on September 2006 from http://www.itu.int/wsis/.