PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 55
Informatyka Ekonomiczna 13 2009

Ilona Pawełoszek-Korek
Technical University of Czestochowa

EXPLOITING THE POTENTIAL OF HIDDEN WEB FOR COMPETITIVE ADVANTAGE

Summary: Most of the approaches described in scientific literature consider the notion of Hidden Web as a challenge for scholars and developers looking for hidden information and working out the applications for crawling Hidden Web. The approach presented here shows the issue of Hidden Web in different way – from the point of view of an e-business entrepreneur.

The objective of this paper is to explore the theoretical and practical issues associated with gaining competitive advantage for Web portals by exploiting an e-business model using Hidden Web resources.

Key words: Hidden Web, competitive advantage, e-business model.

1. Preface – Internet as an environment of e-business

The largest and the most commonly accessible information resource nowadays is undoubtedly the World Wide Web. It is also a business environment for many retailers, entrepreneurs and the place for cultivating social relations among individuals.

Entrepreneurship of the Internet Age poses new challenges, all the actors on virtual arena have similar chances, access to the same media, services and technologies. Therefore enterprises need new models to help them to compete. New strategies are based on professional knowledge, research and nontrivial ideas how to exploit the legacy technology in new, more innovative way.

Most of the approaches described in scientific literature consider the notion of Hidden Web as a challenge for scholars and developers looking for hidden information and working out the applications for crawling Hidden Web. The research activities concerning Hidden Web are mainly conducted around the desire to produce a technological solution to a problem of applying various Artificial Intelligence techniques that can support or automate the process of discovering Hidden Web documents [LDGL04; GRIS03; RAGA01].

The approach presented in the paper hereby shows the issue of Hidden Web in different way – from the point of view of an e-business entrepreneur. The author advocates that the phenomenon of Hidden Web can be used for the benefit of an e-business enterprise. The paper also addresses possible problems and threats related to designing e-business Websites for positioning in search engines.

The objective of this paper is to explore the theoretical and practical issues associated with gaining competitive advantage for Web portals by exploiting an e-business model using Hidden Web resources.

2. Elements of competitiveness in cyberspace

This section explores the contemporary issues of competitiveness in the Internet world. The attempt is to define main competitiveness factors for organizations operating in cyberspace. The specific of online business makes the factors presented herein overtake traditional cost factors.

Online presence

A Website should be properly designed in terms of visual and technical aspects. One of the most important issues in e-business applications acceptance is their legibility to a user. Nowadays the Web becomes the main channel of social and professional communication for many groups of people.

E-business Websites are usually targeted to particular audience, but the possibility of content and display customization is still very important thus the individuals in every professional or social community may have different preferences and perception. As the Internet access becomes more and more popular and disseminates to almost every country in the world, the Web is often the only possible tool of work and communication of people with disabilities. In such instances the crucial thing is to provide a sufficient text and graphics legibility, accessibility by keyboard shortcuts (for people with locomotor system dysfunctions) and descriptive labels for menus and graphics elements for weak-sighted people who use voice browsers.

Supporting different devices

Another continuously growing group are mobile Web users (i.e mobile workers and teleworkers, telecommuters and young people using mobile Web access mainly for shopping and entertainment). Compatibility with different devices requires the site to be built in compliance with W3C standards. Using the XML or XHTML standard accompanied with Cascading Style Sheets makes it much easier to fit the content to small screens of cellular phones or Personal Digital Assistants. In some cases, the URL for the mobile site is the same as the desktop site because of device detection [MOLL08, p. 14] that is a function implemented as server-side solution. The device detection consists in routing specific devices to the most suitable version of the Website. In case of unrecognized device the system offers default WML version that can be properly rendered by most mobile browsers.

Customer support

Because of increasing complexity of products and services coupled with recent changes in sales techniques (from traditional to electronic) a very important issue in todays e-business is customer support. This can be realized in many different ways.

Internet technology provides enterprises with many electronic communication channels that can be offered to a customer. Support can be realized as a communication between a human-consultant and a customer. The means of communication can vary starting from Call Centres through e-mail and instant messaging to on-line video-chat. These are direct means of communication There is another possibility to provide a customer with online access to electronic documents containing user manuals, product and services descriptions or answers to frequently asked questions. Online documents are more suitable for large audience than direct means of communication, and exploiting them is relatively cheaper.

Business processes automation

In business-to-business (B2B) relations an important aspect of strengthening competitive advantage and gaining more customers is transforming a customer, an employee and operations experience. It can be achieved by exploiting Service Oriented Architecture, based on standardized XML Web Services (WS). WS allows to streamline the dataflow between parties. Automation makes cooperation easier and less error prone decreasing the time and costs of human operation.

Web Community involvement

The emergence of Web 2.0 trend provides enterprises with new marketing channel built upon community building software. A good example is Nokia Forum [FORU08] that gathers two groups of customers: users of mobile devices and programmers coding for mobile devices. Several ways exist to benefit from communities. For example, the loyalty of customers that participate in a community may increase, and as a result, these customers buy more of an organization's products. Another way is to use communities as a knowledge base for customer services [DASG06, p. 10]. Nokia Forum is a resource of high quality content and software available free of charge only to registered users and it is also a place for a discussion and sharing knowledge among its participants.

Website positioning

For an e-business enterprise a search engine is undoubtedly a kind of a business partner because a placement of the e-business Website in search results can give it competitive advantage. If products and services can be easily found in cyberspace the Website traffic increases. But it would be misleading to assume that Website placement in search results is the only factor that turns Website visitors into customers. Usually Webmasters care for the Websites to be fully indexed, so the customers can easily find prices and product descriptions. There are many techniques referred to as "Search Engine Optimization" – SEO that aim to increase the number of visitors to a Website. If we continue along this line of thought, the existence of unindexed Web resources (so called Hidden Web pages) is generally unwanted case. However, there are some business models that take advantage of having their Web resources not indexed, and moreover their owners protect them against being accessed directly from search engine results list.

3. Defining an e-business model

There are many definitions explaining what business models are that take into account different aspects of this concept. A business model is the theoretical design of an organization that describes how it makes money on a sustainable basis and how it grows. Business models take on many forms, including [CUCO05, p. 220]:

– straightforward industry classifications,
– methods for trading,
– structural definitions such as functional responsibilities.

The main difference between classical business model and the one for e-business is that the latter describes how the enterprise uses Internet infrastructure and the Web to achieve its goals. Many authors claim that there is no single comprehensive taxonomy for classifying e-business models, yet each of them proposes his own classification based on different criteria.

The classical taxonomy distinguises 12 e-business models [NOWI06, p. 258]: e-shop, e-procurement, e-auction, e-mall, third party marketplace, full service provider, virtual communities, value chain service providers, value chain integrator, collaboration platforms, information broker, trust services. The aforementioned business models for e-commerce and e-business consider the methods of trading and responsibilities.

We define an e-business model as: a description of the roles and relationships among firm's consumers, customers, allies, and suppliers that identifies the major flows of product, information and money, and the major benefits to participants [WEVI01, p. 34].

Defining a business model of the enterprise requires answering six fundamental questions [CHES06, p. 109]:

1. Who are the users/customers to whom the offering and its purpose are useful?
2. What is the value created for users by the offering?
3. What is the structure of the value chain required by the firm to create and distribute the offering (including complementary assets needed to support the firm’s position in this chain)?
4. What are the revenue generation mechanisms for the firm?
5. What is the position of the firm within the value network, linking suppliers and customers (including third –party software developers) and competition?
6. What is the competitive strategy by which the innovating firm will gain and hold an advantage over rivals?

A business model is not the same as strategy. Business models are concerned only with the underlying business concept – they do not take account of competition. When a number of firms adopt a similar business model, the critical determinant of success is which firm will be most successful in deploying its unique attributes in order to create a competitive advantage.

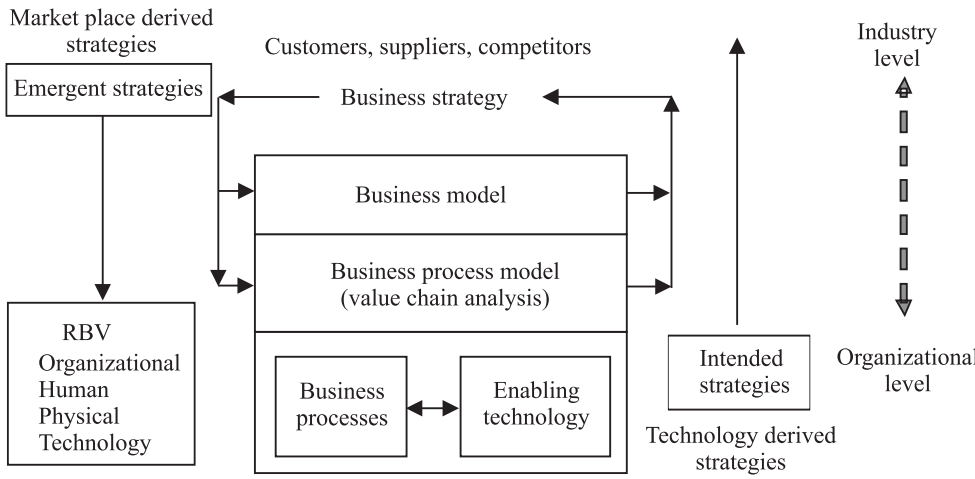

Fig. 1. Business model strategic conceptual framework

Source: [CURR03, p. 50].

A business model is not the same as strategy. Business models are concerned only with the underlying business concept – they do not take account of competition. When a number of firms adopt a similar business model, the critical determinant of success is which firm will be most successful in deploying its unique attributes in order to create a competitive advantage [GRAN05, p. 16]. Business model is a compound concept which encompasses and shows the relations between: strategy, the firm's know-how, enabling technologies and assets. Figure 1 presents a strategic conceptual framework of business model. Every business considering a Website for e-business purposes must develop an underlying strategy based upon well-considered business model. Business models are essential for converting ideas and technologies into economic value [CHES06, p. 107].

The next section provides an introduction to the notion of Hidden Web and explains some reasons for its existance.

4. The notion of Hidden Web

Traditionally getting information from Internet requires using browsers and search engines to locate the information and then accessing or buying it. But it would be a mistake to assume that every information on the Web can be obtained one of these ways. There exist plenty of documents on the Web that are not reachable by search engines crawlers. Consequently these Websites are not indexed and cannot be found in the search results. This unindexed part of Internet is called Hidden Web, Invisible Web or Deep Web.

The information on Hidden Web can relate to every area of human activates including business, science, culture, security and crime counteraction. Thus if found, the hidden information, can be invaluable support for business decision makers, researchers, and any person that engages in activities to acquire domain knowledge.

There are manifold immediate causes of Hidden Web existence:

– a Website is protected by a Webmaster against being indexed,
– a Website cannot be indexed due to its construction and some techniques used by its Webmaster,
– a Website has dynamically created content,
– a Website is not linked to other already indexed Websites and nobody added its URL to search engine database,
– a Website content is only available after filling a form, entering password etc.

Research studies show that Hidden Web is not a marginal matter. The unindexed part of Internet is estimated to be 400-500 times bigger than the easily searchable Web content. Figure 2 shows the pictorial representation of the Hidden Web concept.

Attempts to estimate the quantity of indexed Web pages revealed the number of at least 1.87 billion pages. The approximate minimal size of the indexed World Wide Web is based on the estimations of the numbers of pages indexed by Google, Windows Live Search (Msn Search), Yahoo Search and Ask. From the sum of these estimations, an estimated overlap between these search engines is subtracted [WORL08].

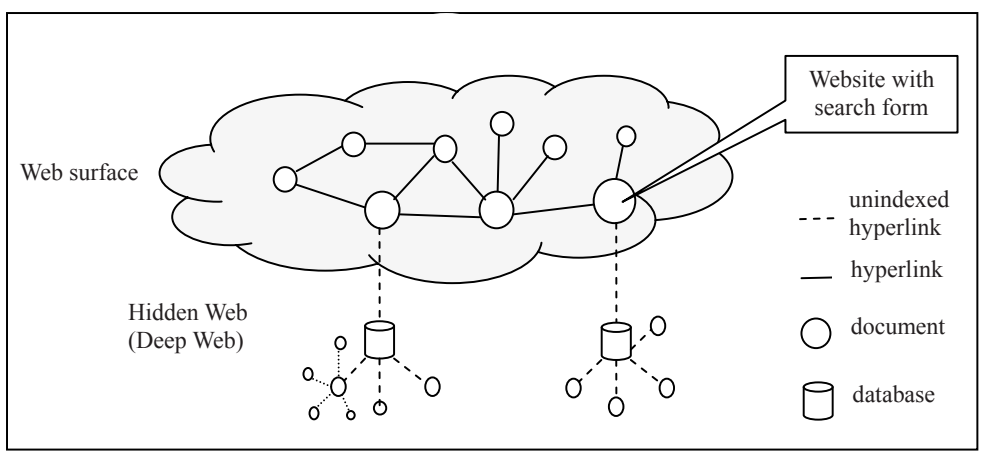

Fig. 2. The concept of Hidden Web

Source: [PAWE08, p. 268].

The vast majority of search engines use so called spiders. These ingenious software programs have only one task – to crawl the Web 24 hours a day, finding and indexing Web pages. These spiders (also called “bots” or “crawlers”) visit a Web

page, read it, and then follow links to other pages within the site. The spider revisits Websites on a regular basis (e.g. every month or two) to look for changes [REYN04, p. 234].

The process of indexing consists in putting the whole content of the page visited by spider to a large database (so called index). When a user enters a query, the search engine checks index database and returns a result from it. The result is a hyperlink and a piece of indexed text taken from a copy of a Website stored in search engine database. The index should be updated every time when the contents of the Website changes, but the system is far from perfect and does not work in a real-time manner. The time needed to re-index takes from a few days to a few weeks. Thus it often happens the query result is out-of-date, and does not match the current Website content.

In order for the Web resource to get indexed, the search engine needs to "know" its location on the Web (URL). The URL can be added to search engine in two ways. The first possibility is entering the page URL manually by a user on the search engine catalog Website. The other way to have a Website indexed is to put a hyperlink to it in the content of some other resource that has already been indexed. Then a spider follows the links on a site and automatically feeds it to search engine database.

The spiders do not execute JavaScript or Ajax code, so consequently they also cannot see the results of the script execution. The spiders can just parse the script, that means they can try to extract some text from the script and index that text. Crawlers can have either rule-based interpretation or try any possible combination of path and file names they find in the script [MASA06, p. 118]. Therefore if the data on a Website is to be displayed as a result of a JavaScript or Ajax code execution it may not be indexed.

Some resources on the Web are intentionally hidden from search engines by the Websites owners. The main problem of e-business strategy is how to attract potential customers to visit the firm's Website, and than how to retain customers longer. Section 5 presents a model for using Hidden Web documents as a marketing tool.

5. E-business model deriving from the concept of Hidden Web

A business model described herein is a compound concept which derives from content provision and it can be referred to as intermediary content provider. For better insight in practical aspects of the enterprise it will be presented as a case study of a Web Portal X. Its main idea is based on contextual targeting of a marketing offer. The model herein is presented according to a framework for defining business models introduced earlier in section 3.

The customers of Portal X

The offering of Portal X is targeted at the group of individual professionals and organizations selling "knowledge intensive products" of information technology do-

main for example: various enterprise information systems, mobile technology, knowledge management systems, security systems, content management systems, data bases and warehouses etc. The clients want to advertise their goods and services by the means of WWW, popularize general knowledge about them. The users need professional, effective marketing.

The value created for users by the offering

It is undeniable that acquiring knowledge about market specifics can provide measurable benefits. The customers of Portal X get a database of individuals and firms potentially interested in their offer. They have a possibility to advertise their products to precisely targeted audience.

The structure of the value chain

The portal needs information about the customer's activities and offerings in a form of business whitepapers, research reports, graphics, presentations, podcasts, videocasts, etc. The information should include:

– general knowledge about the branch and technology offered,
– the problems that can be resolved by means of the given technology,
– what organizational changes are needed to fully exploit the technology in enterprises,
– how those changes reflect the new directions and trends in management,
– some successful examples of deploying the technology,
– what economic and utilitarian value can the technology bring.

The portal needs an editor (or a group of them) to prepare professional marketing materials. The materials are put into a database accessible by the WWW interface only for registered users. The registration is free of charge, it requires only filling out a detailed questionnaire including fields to enter personal and professional data about a person who wants to get access to a given document and some information about her company (like information technology budget, investments planned, the role of a person in a decision making process). Figure 3 presents the example screenshot of a questionnaire.

One of the main competitive elements of every e-business Website is its position in search engine results list. If the materials are Hidden Web resources there is no way for them to by found by Google search or other search engine. Despite that the Portal X Website is high ranked by most search engines. Its competitive strength is a very comprehensive encyclopedia of terms related to information technology. The encyclopedia is accessible freely on the Portal X Website and it is fully indexed and highly ranked by search engines. People looking for terms somehow related to the materials offered by Portal X can easily find their definitions either by search engine or by browsing the Portal itself.

Below every definition there is a list of hyperlinks and summaries of Hidden Web resources accessible after registering. Figure 4 presents an example definition for the term "Data quality" and the links to some whitepapers thematically related to it.

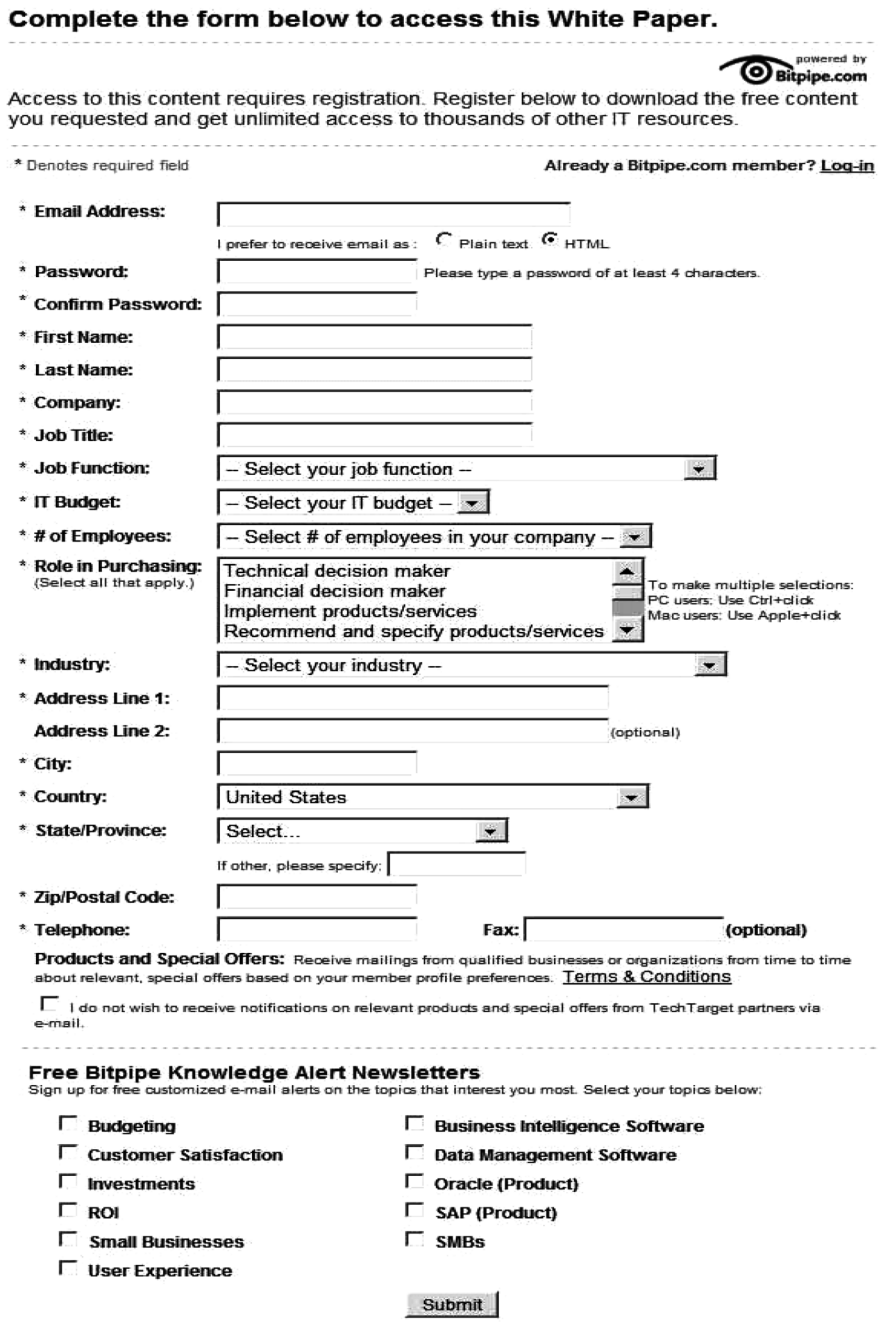

Fig. 3. Example questionnaire

Source: [SEMI08].

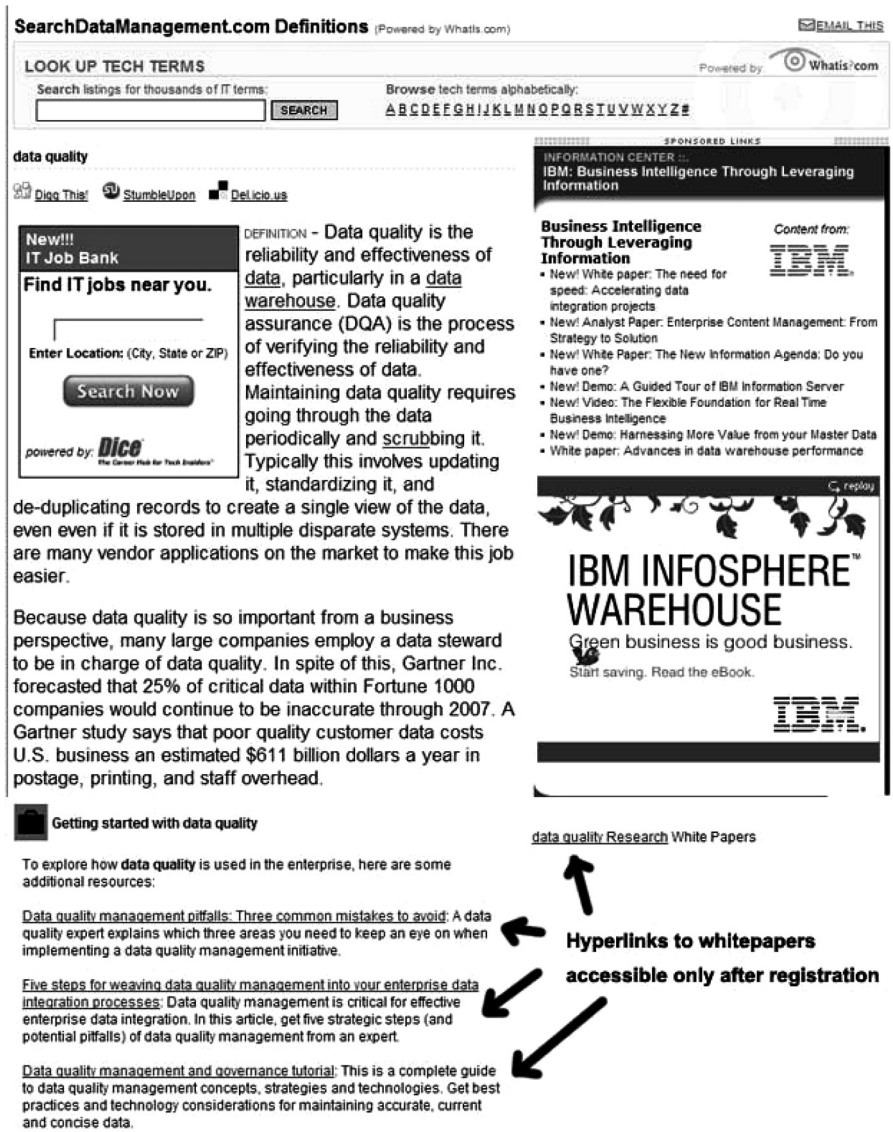

Fig. 4. Example encyclopedia definition with related links to Hidden Web content

Source: [SEAR08].

Even if there are no business whitepapers in a database related to a given definition it is still worth keeping and maintaining it in an online encyclopedia. Every high ranked page may be potentially interested for some advertisers who will pay for placing commercial text links or images leading to external Websites of the informa-

tion technology domain. On the example page on the right side of figure 4 there are advertisements of IBM.

The revenue generation mechanisms

The customers pay for a professional edition of their marketing materials and putting it to a Portal X database. The payment also depends on the number of visitors who got access to the given materials. There is an agreed price per registered visitor. The firms get visitors data and are allowed to send them further commercial information, so they have a database of potential clients with their detailed characteristics.

The position of Portal X within the value network

The role of Portal X is to link suppliers with their potential customers. So the Portal is an intermediary. The suppliers are companies offering information technology products and services. The potential customers are Internet users interested in information technology theory and practice. The parties that contribute in a value chain are also developers providing a software platform for Portal X and Webmasters maintaining underlying databases of the online IT (Information Technology) encyclopedia, marketing materials provided by customers and user accounts. Another important participant is a SEO specialist, who cares about the Portal position in search engine results. As a competition for Portal X we can consider online marketing agencies, consultants and other firms exploiting the same business model.

Competitive strategy

The aim of competitive strategy is to gain main information technology providers to use the services of Portal X. Advertising money will flow to the Portal with the highest volume of users and the highest position in search engine results. Therefore the crucial matters are the maintenance and a continuous improvement of the online IT encyclopedia of Portal X in terms of its functionality and comprehensiveness. Facing the continuous changes in search engines algorithms Portal X should invest in effective SEO activities. The next important thing is improving quality, functionality and attractiveness of the resources available to registered users in terms of multimedia usage, rich content, easily accessible and effective search mechanisms. The pages of Portal X should be adjusted to meet the demands of growing number of mobile Internet users.

6. Conclusions

The advent of Internet and diverse e-business models produced significant changes to entrepreneurial practices nevertheless competitiveness and innovativeness are still key factors of sustainable economic growth. For every e-business it is essential to have a broad understanding of online customer needs, as well as market trends and behaviour. Every e-business model can be potentially useful, but the key for success

is to recognize a chance for product differentiation and providing new value able to strengthen competitive forces of an enterprise. The business model presented in hereby paper is an innovative variety of content provider model. The example of Portal X presented hereby is focused on IT products, but the model can be as well applied to many different fields (like: medical supply, construction industry, heavy industry, transport, etc.). It is applicable especially in advertising knowledge intensive services and commodities meant for institutional customers.

References

[CHES06] Chesbrough H., *Open Business Models: How to Thrive in the New Innovation Landscape*, Harvard Business School Press, Boston 2006.

[CUCO05] Curtis G., Cobham D., *Business Information Systems analysis, design and practice*, Prentice Hall, Edinburgh Gate, Harlow, Essen, England, 2005.

[CURR03] Currie W. (ed.), *Value Creation from E-Business Models*, Butterworth-Heinemann 2003.

[DASG06] Dasgupta S., *Encyclopedia of Virtual Communities and Technologies*, Hershey, PA: Idea Group Reference, 2006.

[GRAN05] Grant R.M., *Contemporary Strategy Analysis: Concepts, Techniques, Applications*, Blackwell Publishing 2005.

[GRIS03] Gravano L., Ipeirotis P.G., Sahami M., *Qprober: A system for automatic classification of Hidden-Web databases*, ACM Transactions on Information Systems 2003, 21(1).

[FORU08] http://www.forum.nokia.com, 16.09.2008.

[LDGL04] Lage J.P., Da Silva A.S., Golgher P.B., Laender A.H., *Automatic generation of agents for collecting Hidden Web pages for data extraction*, Data & Knowledge Engineering 49, 2 May. 2004.

[MASA06] Masanès J. (ed.), Web Archiving, Springer 2006.

[MOLL08] Moll C., *Mobile Web Design p. 14*, www.w3.org/2006/07/Mobile_Web_Design.pdf, online: 16.09.2008.

[NOWI06] Nowicki A., *Komputerowe wspomaganie biznesu*, Placet, Warszawa 2006.

[PAWE08] Paweloszek-Korek I., *Methods of discovering Hidden Web resources*, [in:] *Informatyka dla przyszłości*, ed. J. Kisielnicki, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa 2008.

[RAGA01] Raghavan S., Garcia-Molina H., *Crawling the Hidden Web*, Proceedings of the 27th International Conference on Very Large Data Bases, San Francisco, CA, 2001.

[REYN04] Reynolds J., *The Complete E-Commerce Book: Design, Build & Maintain a Successful Web-based Business*, CMP Books, 2004.

[SEMI08] http://searchcio-midmarket.bitpipe.com/, 16.09.2008.

[SEAR08] http://searchdatamanagement.techtarget.com/, 16.09.2008.

[WEVI01] Weill P., Vitale M.R., *Place to Space: Migrating to Ebusiness Models*, Harvard Business School Press, Boston, Massachusetts 2001.

[WORL08] http://www.worldwidewebsize.com/, online: March 2008.

WYKORZYSTANIE POTENCJAŁU UKRYTEJ SIECI WEB DO OSIĄGNIĘCIA PRZEWAGI KONKURENCYJNEJ

Streszczenie: Większość podejść opisywanych w literaturze naukowej rozważa zjawisko głębokiego (ukrytego) Internetu jako wyzwanie dla badaczy szukających ukrytej informacji lub programistów pracujących nad aplikacjami przeszukującymi zasoby niedostępne dla tradycyjnych wyszukiwarek. Podejście zaprezentowane tutaj pokazuje zagadnienie ukrytej sieci Web w odmienny sposób – z punktu widzenia przedsiębiorstwa e-biznesowego. Celem artykułu jest rozważenie teoretycznych i praktycznych aspektów związanych z osiąganiem konkurencyjnej przewagi portali internetowych działających w oparciu o model biznesowy wykorzystujący ukryte zasoby sieci.