韓國 컴퓨터情報學會 論文誌
第12卷 第3號, 2007. 7.

2007-12-3-4-4

키워드 분석을 이용한 개인화 모바일 웹 뉴스 컨텐츠 생성에 관한 연구

한 승 현*, 임 영 환**

A Study on Personalized Mobile Web News Contents Creation using Keyword Analysis

SeugnHyun Han*, YoungHwan Lim**

요 약

본 연구에서는 웹 뉴스 채널 컨텐츠의 키워드 분석을 이용한 개인화된 모바일 웹 컨텐츠 생성 방법에 대해 제안한다. 기존의 웹 사이트의 뉴스기사 검색에서 제공하는 RSS와 연계된 웹 컨텐츠에서 빠르게 데이터를 획득하고, 키워드 분석을 통한 개인화 기법을 적용하여 컨텐츠를 필터링한다. 제안한 방법을 사용함으로써 수많은 뉴스 채널에서 보다 빠르고 쉽게 모바일용 웹 컨텐츠를 생성할 수 있어 컨텐츠 제작비용을 줄일 수 있다. 또한 키워드 분석을 이용하여 무선 인터넷 사용자들의 보다 세밀한 관심영역에 대응할 수 있으며 컨텐츠 필터링과 컨텐츠 접근에 대한 만족도를 향상시킬 수 있다.

Abstract

This research proposes a personalized mobile web contents creation method that uses web news channel contents-based keyword analysis. It promptly acquires data through the RSS and RSS-linked web pages which have been supplied by the existing web sites for a news search. And then It applies a personalization method using keyword analysis in contents filtering and generation. The proposed method will make creating mobile web contents easier while lowering wireless contents production costs. Moreover, It can be improved a user satisfaction for contents filtering and access with using keyword analysis that fits in with a matter of user's specific interest.

▸ Keyword : 모바일 웹(Mobile Web), 컨텐츠 필터링(Contents Filtering), 개인화(Personalization), 트랜스코딩(Transcoding), 키워드 분석(Keyword Analysis)

• 제1저자 : 한승현
• 접수일 : 2007.4.16, 심사일 : 2007.4.25, 심사완료일 : 2007. 4.27.
* 숭실대학교 대학원 컴퓨터학과 ** 숭실대학교 대학원 미디어학과 교수

Ⅰ. Introduction

Web services occupy most of the internet services. Many internet users search and visit web-sites to access and use the information that they need. However, the limited browser installed in wireless devices, the low hardware performance, and the inconvenient interface contribute to the inefficiency in terms of time and cost when it comes to accessing the internet with wireless devices.

Various methods that convert original web-pages to mobile web-pages for the OSMU(One Source Multi-use) of the digital convergence environment have been studied[1, 2]. Although pages are generated relatively promptly, these methods lack the consideration of the users and the services that are optimized for the different devices as well as too many pages being generated. The following issues must be considered in order to complement such shortcomings.

First, the characteristics and preferences of users must be taken into account. In other words, a wireless web service that considers the preferences of users will prevent resources from being wasted while enhancing user satisfaction.

Secondly, wireless pages optimized to a variety of wireless devices should be serviced. Services restricted to specific devices do not consider the different sizes and performances of wireless devices.

Thirdly, the changes made in an original web contents must be automatically reflected in the wireless devices. It is extremely inefficient in terms of time and money when relying on manually reflecting such changes in the wireless devices.

This research proposes a servicing method in which the desired information can be serviced to wireless devices by using concerned areas and concerned keywords. First of all, mobile web contents are generated by classifying the preferred keywords and the related keywords of a user. The suitability and preference of the generated contents are analyzed and this analysis is used to generate personalized web contents. Moreover, by using RSS(RDF Site Summary) that can generate active channels, wireless internet can easily be accessed because new contents suitable for mobile environments can be generated at any time.

The organization of this paper is as follows. Chapter 2 introduces the personalization methods of web contents and related researches. Chapter 3 describes how the personalized channel contents are created and how RSS and HTML data is extracted and converted. Chapter 4 detail s the test results and performance evaluation. Finally, Chapter 5 concludes the research.

Ⅱ. Related Studies

Contents generation and personalization methods for mobile environment, which has more restrictions compared to the PC environment, have actively been researched.

Power Browser[3] provides a function in which existing web contents can be restructured and summarized to a simpler version. If there are many pages that are linked to one page then they are positioned to the above contents.

Real-time data transfer is difficult when manually converting web contents for mobile[4, 5] use because the overhead consumed in the conversion occupies a large amount. All the methods listed above do not consider the individual preferences of users when it comes to contents conversion.

The data in general personalization use may organize explicit data such as the name, address, age, preference, and interests provided by the users and implicit data acquired through the analysis of the flow between the user and the web server and such data analysis are done through database or log files.

Some of the data analyzing methods include reasoning using the association-rule and the CBR (cased-based reasoning)[6, 7]. The latter is based on cases of other users who are not related to a specific user and the former is a method in which it searches

for a rule that is identical to a case for reasoning however it is inefficient in that reasoning becomes impossible when a rule is not detected. However, such method is only possible when a considerable amount of cases are accumulated and it is extremely risky for real-time services to allocate a long time detecting and analyzing the rules in newly introduced data. In addition, the method of simple log analysis has the following shortcomings[8]. First, personalized contents and original web contents may share a faulty search structure. Secondly, it is difficult to identify the contents that a user actually wants because only the page access frequency is counted. Thirdly, it is troublesome to adopt the large-capacity accumulated log data to the real-time handling system if it has to be analyzed through complex handling process.

The following measures can be taken to resolve the above restraints.

First, the original web contents and the converted personalized mobile web contents both provide identical searching structures and a search method in which the access process is minimized to access new keywords is provided. Nevertheless, it is designed that such features do not directly affect the class that contents are located in. Secondly, contents are generated by analyzing the related keywords and keyword access frequency within contents rather than simple page access frequency.

Thirdly, a faster preprocessing through keyword analysis is introduced and applied to real-time contents generation rather than analyzing the large amount of logs.

III. System Organization

3.1 Processing Structure of System

The structure of the system proposed in this research is as shown in Figure 1. When a wireless device requests for service the device manager analyzes the protocol header information and identifies the connected browser type, device type, and supported image and sound. This information is recorded in the personal profile through the profile manager.

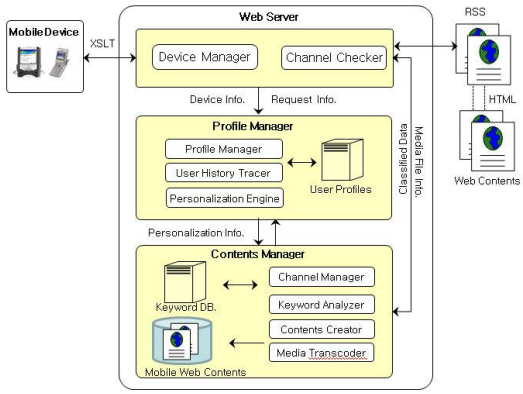

그림 1. 시스템 구조
Fig. 1 System Structure

The profile manager acquires the requested keyword, related keyword, and preferred genre during the browsing and adds it to the personal profile. These data are used in the individual -ization through the personalization engine. The channel manager groups by genre of keyword when a user requests a keyword from a certain genre. The keyword analyzer of the contents manager classifies the relation information among the channel composition keyword. The contents creator acquires the necessary parts in each item recorded in the RSS channel. Full text feeds are extracted from the linked HTML page for generating the XML data to be used in the wireless page generation. In addition, media file information and original data are transferred so that the media transcoder can handle the full-text feeds related images. The media transcoder is a set of modules that converts the size, quality, and format of the images(bmp, jpg, gif, png) used in webs so that they are suitable for mobile devices. Finally, the XML+XSL that is defined specific to mobile devices through XSLT is serviced in the forms of XHTML and WML2.0.

3.2 Applying Personalization Method

In this research, detailed information about concerned areas is obtained rather than from the number of times a page has been accessed. Wireless pages specific to individual areas can be created through generating channels according to acquired keywords. Moreover, the preferred genres and keywords are constantly updated by reacquiring user history because they are always subject to changes. Figure 2 illustrates the reflections made in the personal profile through searching the user preference data and browsing history.

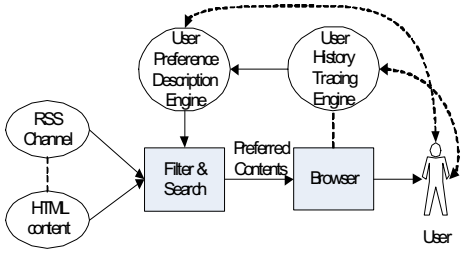

그림 2. 사용자 선호도 반영의 흐름
Fig. 2 Flow of user preference reflection

The description of the preference information of personal profile is assorted according to the data that have been sorted by the structure as shown in Table 1. Such classification information was also used in evaluating the user satisfaction about the personalization method.

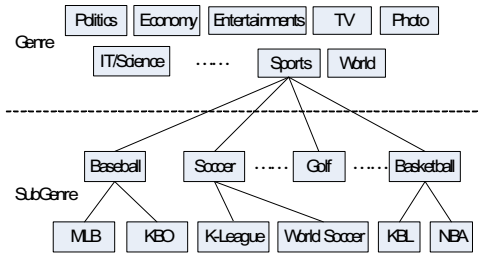

그림 3. 뉴스 장르에 대한 구분
Fig. 3 News genre classification

Using news article-based RSS in this research, the basic genres are organized as shown in Figure 3. Preferred genres and sub-genres are implemented in the personal profile through such genre classification

Figure 4 shows the relationship between the keywords that appear in each genre and the names of the genres and sub-genres. This relationship is an important element when it comes to personalizing the contents to provide to the users.

표 1. 사용자 선호도의 기술을 위한 기본 분류
Table 1. Basic classification according to user preferences

Channel maintenance Preferences	Concerned Preferences		
	Keyword Preferences	Classification Preferences	Personal View Preferences
Publication _source Channel _creation _keyword Channel _connection _info. Channel _update _info. Renewal _items Deleted _items	Subject _keyword Relational _keyword Content _exactness Content _preferences Date period Image _existence	1st_Genre 2nd_Genre 1st_SubGenre 2nd_SubGenre	Personalization_appliance Image _existence Feeds _length

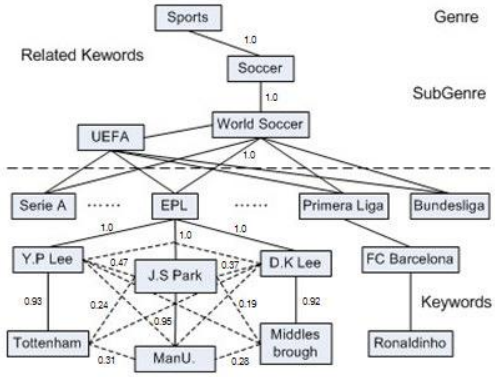

그림 4. 키워드 분류와 키워드 간의 연관계수
Fig. 4 Keyword classification and Related coefficient among keyword

This is because when contents that lack such connection becomes a concerned contents it indicates an error in the personalization method. Therefore the following equations and algorithms were used to examine

the preference and the suitability of the generated contents as well as the connection among keywords.

The coefficient related to an arbitrary keyword m is obtained from the following equation(1).

$$RC(k_m) = \frac{freq(k_m, FTF) + \left(\frac{\sum_j FTF(Rk_j)}{freq(k_m, FTF)}\right)}{\sum_n FTF(Pk_n)}$$

where $(RC(k_m) \le 1)$ ····· (1)

$RC(k_m)$: Relational Coefficient of Keyword m

$freq(k_m, FTF)$: Frequency of Keyword m in the Full Text Feeds

$FTF(Pk)$: Preferred Keyword in the Full Text Feeds

$FTF(Rk)$: Related Keyword in the Full Text Feeds

The following equation(2) is used for the suitability examination of the contents generated according to a requested keyword. Whether or not the contents are correctly created depends on the number of times the corresponding keyword appears in the contents.

$$CSC(k_j) = \frac{freq(k_j, SJ) + freq(k_j, FTF) + \sum_j FTF(Rk_j)}{freq(k_j, FTF)}$$

where $(CSC(k_j) > 1)$ ····· (2)

$CSC(k_j)$: Content Suitability Check of Created Channel by Keyword j

$freq(k_j, SJ)$: Frequency of Keyword j in the Subject

$freq(k_j, FTF)$: Frequency of Keyword j in the Full Text Feeds

$FTF(Rk)$: Related Keyword in the Full Text Feeds

The contents preference of the channel generated by a requested keyword is calculated with the following equation(4). Consequently, whether or not the channel includes the contents that the user is interested in is calculated using all the keywords generated in the channel, the keywords that the user frequently requested lately, and the frequency of the corresponding key of all the full-text-feeds in the channel.

$$FACC(k_i) = freq(k_i, \sum_m FTF_m)$$ ····· (3)

$$UPC(k_i) = \frac{\sum_g CCk_g}{\sum_r RRk_r} + FACC(k_i)$$ ····· (4)

$FACC(k_i)$: Frequency of Keyword i in all Full Text Feeds on a Created Channel by Keyword i

$UPC(K_i)$: User Preference of Created Channel by Keyword i

CCk: Channel Creation Keyword

$RR(k)$: User's Recently Request Keyword

Algorithm 1 shows the concerned channel configuration process that considers user preferences obtained by using the above (1), (2), (3), (4) equations.

알고리즘 1. 사용자 관심 채널의 설정 Algorithm.1 Configuring concerned channel of user
(i) Re-enter the request keyword if it doesn't exist in the keyword list in a certain sub-genre. (ii) The newly entered keyword is added as a related keyword within the corresponding sub-genre. (iii) The keywords that other users requested for the same sub-genre is classified as related keywords. (iv) All full-text-feeds of the created channels are searched and the related coefficients among related keywords are obtained and then applied by equation(1) and the related coefficients data are recorded in the keyword database. (v) All the keywords selected as related

keywords use equation(2) for the verification of contents suitability for all the channels created within a sub-genre.

(vi) The user preferences of all the contents within the channel that passed the contents suitability check as mentioned in the above (v) are calculated according to equation(4) and the preference of the requested keyword (Rk_n) is recorded in the personal profile.

(vii) The sub-genre that contains the requested keyword (Rk_n) with high user preference is given the highest priority in terms of location and such information is also recorded in the personal profile.

3.2 Applying the Personalization Method

The contents of the channels that are created according to requested keywords must be deleted and regenerated every now and then in order to maintain the real-time information servicing. The updates of the requested keywords are confirmed and handled as shown in Figure 5 to receive the most recent information. Through this process, users can always access updated information with mobile web pages. Moreover, the existence of updates can be handled more promptly than other contents sources because RSS is used as the original contents source.

Whether or not a channel is already generated with the corresponding keyword is determined when a request is received. If it has not been generated, a wireless web channel contents is generated by obtaining the data from the RSS source channel. Here, the inappropriate contents within the channel are filtered with equation(2) during the contents filtering process. If a generated channel exists, the original channel is examined whether or not it contains added items. If there are added items, the wireless web channel contents is regenerated otherwise the wireless web channel contents that is personalized for user views is sent to the server's response.

IV. Test Results

4.1 Test Environment

The test environment is as follows. Windows2003 Standard Edition O/S, web server IIS 6.0, Pentium-4 3.0GHz CPU, 1GB memory, and Visual Studio 6.0 and ASP were used as the development tools. OpenWaveV.7 Simulator and LG-SD910 mobile phones were used for the client connection test.

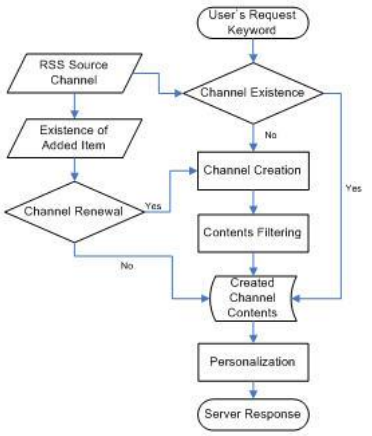

그림 5. 채널 컨텐츠의 생성 및 갱신 알고리즘
Fig. 5 The channel contents generation and updating algorithm

4.2 Filtered Contents Generation and Access

Figure 6 is a chart that shows the capacities of the data that composes each channel and the averaged time spent in contents connection, personalization method, channel generation, contents filtering, and the channel changing verification of the original web for the six sample keywords extracted from the twenty requested keywords obtained from the user's preferred genre. Although the creation time took longer than the channel checkup time or personalization time, a real-time service through wireless internet was possible as the result of the test. This is due to the fact that images are minimized through the image transcoding process and mobile web-pages are created by our keyword-based contents filtering.

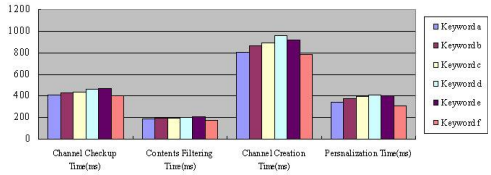

그림 6. 사용자의 관심 키워드에 의한 채널 생성의 결과 (ms)
Fig. 6. The result of the Channel Creation by User's Concern Keywords (ms)

Figure 7 shows average access time of mobile phone web-browser that is in use CDMA-EVDO networks for the created mobile news contents(a~f) by the sample keywords(a~f) versus total service preparation times. A real-time service through wireless internet is possible because total service preparation times are much lower than mobile browser's access times for each channel contents.

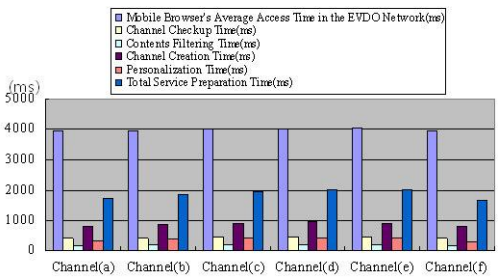

그림 7. 생성된 컨텐츠에 대한 모바일 브라우저의 평균 접근 시간과 토탈 서비스 준비 시간의 비교
Fig. 7 Mobile browser's average access time for the created contents versus total service preparation time

The genre classification, related keyword, and user preference of corresponding request keyword based on the classification rule for an arbitrary keyword(m) among the request keywords within identical sub-genres that an arbitrary user(i) created are shown in Table 2. For example, the related coefficients of Keyword(0) and Keyword(5) both turned out to be 0.47 whereas no identical coefficients were found between Keyword(0) and the rest of the keywords and between Keyword(5) and the rest of the keywords.

표 2. 임의의 요청 키워드와 연관 키워드간의 연관계수 및 사용자 선호도 분석
Table. 2 Analysis of related coefficients and user preferences between arbitrary request keywords and related keywords

Genre (g)	Sub Genre (s)	Request Keyword (m)	Related Keyword (n)	Related Co-efficient	Preference of Request Keyword (m)
Sports	Soccer /World Soccer / EPL	Keyword (0)	Keyword(1) Keyword(2) Keyword(3) Keyword(4) Keyword(5) Keyword(6) Keyword(7) Keyword(8)	0.24 0.95 0.69 0.56 0.47 0.24 0.19 0.37	5.86
Sports	Soccer /World Soccer / EPL	Keyword (5)	Keyword(0) Keyword(1) Keyword(2) Keyword(3) Keyword(4) Keyword(5)	0.47 0.93 0.44 0.38 0.41 0.62	4.87

The user preference of the user's request Keyword(m) has the same result as the one where equation(4) is applied to obtain the contents preference of a created channel because the system's channel creation is based on the requested keyword.

The following Table 3 shows how grouping through the personalization according to keyword preferences within an arbitrary sub-genre (Sports-Soccer-World Soccer) is organized.

The genres are categorized within the Meta data of the XSL documents that are applied to the corresponding genres because all the sub-contents that are created by an arbitrary keyword of a single sub-genre have identical genre classification.

It is apparent that the request frequency of each of all the user's keywords does not coincide with the keywords of each specific individual. Grouping was applied to the keywords that have values over the threshold calculated according to the personal preference and the font size and background color was configured to appear differently. The application of

grouping through personal preference is shown in (c) of Figure 8.

표 3. 임의의 서브장르내의 키워드의 선호도에 따른 그룹화 적용
Table 3. Grouping according to the preference of keywords within an arbitrary sub-genre

News Genre (Sample)	Request Keyword	Request Frequency	Personal Preference	Grouping By Freq.	Font Size	Back-ground
스포츠-축구-해외축구 (Sports-Soccer-World Soccer-EPL)	박지성	212	5.86	group 1	15	#FFFF00
	맨체스터	126	3.93	group 1	13	#CCFFCC
	호나우두	114	4.28	group 2	15	#FFFF00
	볼튼전	135	3.71	group 2	13	#CCFFCC
	이영표	143	3.87	group 1	13	#CCFFCC
	루니	47	2.79	group 1	11	#CCCC66
	이동국	97	3.45	group 2	13	#CCFFCC
	FA컵	53	2.12	group 2	11	#CCCC66
	토트넘	76	2.87	group 2	11	#CCCC66
	미들즈브러	68	2.96	group 2	11	#CCCC66
	블랙번	75	3.12	group 3	13	#CCFFCC
	첼시	82	3.01	group 3	13	#CCFFCC
	C.호나우두	67	0.83	group 2	9	#FFCC66
	레딩	46	0.94	group 2	9	#FFCC66
	설기현	52	1.75	group 3	10	#FFCCCC
	……	……	……	group 3 group 3 group 2 group 2 group 5 group 5 group 4 ……	……	……

Figure 8 displays the mobile web contents browsing generated through the method proposed in this research.

(a)~(e) of Figure 8 show the browsing of the personalized mobile web contents, (f)~(g) show advanced option configuration, and (h) shows the entering of the desired keyword if it is not found in the sub-genres. A new channel keyword which is not exists as an enumerated keyword list becomes the related keyword and user preferred keyword within the relevant channel

그림 8. (a)장르의 선택, (b)서브장르의 선택, (c)개인화가 적용된 채널 키워드의 선택, (d)선택한 채널 키워드에 대한 컨텐츠, (e)이미지가 포함된 Full Text Feeds의 탐색, (f)이미지 선호 옵션을 제거한 경우, (g)개인화가 제거된 채널 키워드의 선택, (h)추가적인 채널 키워드의 입력 화면
Fig. 8 (a) genre selection, (b) sub-genre selection, (c) personalized channel keyword selection, (d) contents of the selected channel keyword, (e) browsing the image-included full-text-feeds, (f) image preference option disabled, (g) selection of channel keyword without personalization, (h) additional keyword entering screen for nonexistent channel keyword

Ⅴ. Conclusion

Although the universalization of devices that support faster wireless connections along with being able to use the internet regardless of the location, limited browser capabilities, large load of data of existing internet pages, relatively expensive web connections, and the lack of useful contents were the obstacles when using mobile web services. This research proposes a method in which real-time contents that reflect user preferences can be accessed while diminishing the mobile web access difficulties through a creative personalization method. Moreover, active information

updating is made possible through RSS servicing. Successful results were achieved without performing personalization based on complex reasoning and simple frequency calculation through the proposed creation process, channel generation based on keywords and related keyword analysis. The personalized mobile web contents creation using keyword analysis can be effective enough to service a wireless web news contents. The strength of the proposed system is the fact that services can be optimally personalized for users and wireless devices and that it can be applied to all contents that support RSS servicing.

Acknowledgement

본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.

참고문헌

〔1〕 W.W.Lu "Compact multidimensional broadband wireless: the convergence of wireless mobile and access, (Journal) IEEE Communications Magazine, Vol.38, No.11, pp.119- 123, 2000.

〔2〕 V. Kwan, F. Lau, C. Wang, "Functionality Adaptation: A Context Aware Service Code Adaptation for Pervasive Computing Environments", IEEE/WIC International Conference on Web Intelligence, Halifax, Canada, October 13-17, 2003.

〔3〕 O. Buyukkokten, H. Garcia Molina, A. Paepcke and T. Winograd, "Power Browser : Efficient Web Browsing for PDAs", Proceedings of the Conference on Human Factors in Computing Systems, pp.430-437, 2000.

〔4〕 http://www.avantgo.com

〔5〕 http://www.earthlink.net

〔6〕 R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkanmo, "A Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining, ed. U. Fayyad et al., AAAI Press: Menlo Park, CA, pp.307-328, 1996.

〔7〕 P. Domingos, "Unifying Instance-Based and Rule-Based Induction, Journal of Machine Learning, Vol.24, No.2, pp.141-168, 1996.

〔8〕 J. Boulicaut, A. Bykowski and C. Rigotti, "Free-sets: a condensed representation of Boolean data for the approximation of frequency queries", Data Mining and Knowledge Discovery, Vol.7, pp.5-22, 2003.

〔9〕 송특섭 외 6인, "MyNews: 모바일 환경에서 사용자 관심사를 고려한 XML문서 트랜스코딩", 정보처리학회논문지, 제12-B권 제2호, pp.181-190, 2005.

〔10〕 전영효, 황인준, "모바일 사용자를 위한 웹 서비스 페이지 개인화 기법", 정보과학회논문지, 컴퓨팅의 실제 제11권 제1호, pp69-80, 2005.

〔11〕 Munchurl Kim, et al, "Agent-Based Intellignet Multimedia Broadcasting within MPEG-21 Multimedia Framework", ETRI Journal, Vol. 26, No.2, pp136-148, 2004

〔12〕 Rehm, G. "Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage", System Sciences, HICSS. pp.1143- 1152, 2002.

저 자 소 개

한 승 현

2002년: 숭실대학교 대학원 컴퓨터학과 공학 석사

2003년~현재 : 숭실대학교 컴퓨터학과 박사과정

관심분야 : 모바일, 멀티미디어 시스템, 유빅쿼터스 컴퓨팅

임 영 환

1977년: 경북대학교 수학과 이학사

1979년: 한국과학기술원 전산학과 이학 석사

1985년:Northwestern Univ. 전산학과 이학 박사.

현. 숭실대학교 미디어학과 교수

관심분야 : 멀티미디어, 트랜스코딩