WORKING DRAFT

International Conference on Universal Digital Library
2007 PROGRAMS THAT ARE OPEN TO THE PUBLIC

Friday, November 2, Posner Center, Carnegie Mellon University

1:30 p.m. Greetings from university and delegation dignitaries

Dr. Raj Reddy, Carnegie Mellon University
Ms. Li Xiaoming, Ministry of Education, China
Prof. Yunhe Pan, China (video)
Dr. N. Balakrishnan, Indian Institute of Science

2:30 p.m. Forum on International Initiatives on Copyright
Open to campus community
Webcast

Dr. John Mark Ockerbloom, University of Pennsylvania *moderator*
Dr. Michael Shamos, Carnegie Mellon University
Dr. Zhipan Wu, Peking University (video)
Dr. N. Balakrishnan, Indian Institute of Science
Dr. Ismail Serageldin, Library of Alexandria

4 p.m. Break

4:30 p.m. Intellectual Property Rights in the Digital Age: Reflections on Why Copyright Law Must Change
Open to campus community
Webcast

Dr. Ismail Serageldin, Library of Alexandria

Saturday, November 3, Newell-Simon Hall, Carnegie Mellon University

9 a.m. Dr. Robert Kahn, Corporation for National Research Initiatives (CNRI)

9:45 a.m. Dr. Michael Lesk, Rutgers University

10 a.m. Break refreshments available

10:30 a.m. Dr. Wen Gao, Peking University, China

11 a.m. Dr. Yueting Zhuang, Zhejiang University

11: 30 a.m. Post-Digitized Books: The Future

Dr. Raj Reddy, Carnegie Mellon University
Katy Borner, Indiana University
Dr. Jaime Carbonell, Carnegie Mellon University
Stephen Griffin, National Science Foundation

1:30 p.m. Parallel Sessions of Contributed Papers Locations/Moderators [topics]

Session 1:
Legal, Commercial
Licensing Data to Keep it Open
AUTHOR Richard Wallis, Talis Information Ltd., United Kingdom
ABSTRACT Many are under the misapprehension that a Creative Commons Logo, or a "this data is public domain" message will protect data published on the web. There are two types of data - creative works, which can be protected by copyright; and factual data that cannot be protected by copyright. Both types of data need to be considered when openly publishing data on the web, if the initial open intentions are to be preserved and protected.

Open data initiatives want and need a community to build around them. The altruistic inclinations of such communities is what helps to build the value of the initiative for all. That altruism is only fostered if the community trusts those hosting the data, and that their contributions and efforts will not eventually be hijacked by closed commercial interests.

Much library metadata are facts and therefore not copyrightable so how can an open database, of Marc records for instance, be preserved and protected from such concerns? In Europe an open licensing of Database Right may suffice, but beyond European boarders the solution has not been so clear. Creative Commons and Open Software licenses, such as GPL, are often quoted on sites, but as they depend on copyright they are no protection for factual data. These, often useless, assertions are usually advertised to protect those who host the

data from unscrupulous people that may wrongly copy and use it. It is equally important that the contributors of data are protected from those that host it.

This session will explore some of the detail behind these issues, and look at recent developments on the licensing front which address these concerns, stimulated by the open contribution of draft licenses by Talis.

Human Factors, Legal, Policy

Faculty Self-Archiving: The Gap Between Opportunity and Practice

AUTHOR Denise Troll Covey, Carnegie Mellon University, United States

ABSTRACT Carnegie Mellon's vision of the "universal digital library" is free-to-read access to the cultural and intellectual heritage of humankind, including traditional and innovative scholarly work. The success of initiatives to incorporate free-to-read versions of scholarly publications in the digital library hinges on the participation of faculty authors. Authors must retain the right to self-archive their work or publish in open access journals. Despite the availability of publisher self-archiving policies and substantial investments in software and protocol development, educational archiving remains quite low—so low that the National Institutes of Health has requested a mandatory policy to require self-archiving of the work it funds.

To improve our understanding of the self-archiving practice of campus faculty and the opportunity to self-archive in different disciplines, Carnegie Mellon University Libraries is conducting a study of faculty publication lists available on the web. In phase I of the study, publication lists are being analyzed to identify publication type and access type. In phase II of the study, all of the journal publications are being analyzed to further determine whether the work was, or could have been self-archived in compliance with publisher policy. To date, Phase I and II of the study have been completed for the departments in Carnegie Institute of Technology (CIT). Phase I has been completed for the School of Computer Science.

The findings for CIT are enlightening. Only a third of the technical reports and conference papers have been self-archived. Few books or book chapters have been self-archived. Only 24% of the journal articles have been self-archived, and at least a third of those breach publisher policy. Yet publisher policy would allow self-archiving of 67% of the journal articles published by CIT faculty.

Publisher policy and faculty behavior of course vary across disciplines. There appears to be no correlation between the opportunity to self-archive in a given discipline and faculty practice. For example, 77% of the journals in which faculty in the department of Engineering and Public Policy (EPP) publish allow self-archiving, but EPP faculty have self-archived only 15% of their articles. The data suggest that faculty either do or do not self-archive, and that many do not know or do not care about the subtle nuances of publisher self-archiving policies regarding the version that can be archived, required text or links to the

publisher's website, etc. It is likely that the variation in publisher policies and the requirements to change the descriptive text or online version after publication are too complicated to encourage or secure total compliance.

The data from the study will be used to inform faculty of the opportunity to self-archive in their discipline and to spark discussion of why and how to archive. The University Libraries aims to work with faculty to remove barriers and to close the gap between opportunity and practice.

Legal

Digitization and the Copyrighted Public Domain

AUTHOR Jason Mazzone, Brooklyn Law School, United States

ABSTRACT This paper examines how false copyright claims over public domain works impede mass digitization. Though anybody can digitize (or otherwise copy) a work that is in the public domain, many publishers of text versions of public domain works assert (falsely) copyright in these works, and work vigorously to prevent unauthorized copying. In the United States, falsely asserting a copyright is a criminal offense but prosecutions are extremely rare (zero in recent years) and there exists no civil penalty for false copyright claims. Publishers therefore have an incentive to assert copyright wherever they can, even over public domain works, and to collect fees for copies and otherwise limit distribution. Digital libraries and other potential users of public domain works are deterred by false copyright claims backed up by the threat of lawsuits. The paper proposes several ways for digital libraries to take account of these risks and to confront the problem of a copyrighted public domain. It also explores the responsibility of digital libraries to provide accurate copyright information about their own products and collections, and the ways in which accuracy serves the libraries' own interests. In its analysis, the paper draws lessons from abroad.

Human Factors, Policy, Technical

The Biodiversity Heritage Library

AUTHOR Thomas Garnett, Smithsonian Institution Libraries, United States
(presented by Catherine N. Norton, Smithsonian Institution)

ABSTRACT This paper examines how false copyright claims over public domain works impede mass digitization. Though anybody can digitize (or otherwise copy) a work that is in the public domain, many publishers of text versions of public domain works assert (falsely) copyright in these works, and work vigorously to prevent unauthorized copying. In the United States, falsely asserting a copyright is a criminal offense but prosecutions are extremely rare (zero in recent years) and there exists no civil penalty for false copyright claims. Publishers therefore have an incentive to assert copyright wherever they can, even over public domain works, and to collect fees for copies and otherwise limit distribution. Digital libraries and other potential users of public domain works are deterred by false copyright claims backed up by the threat of lawsuits. The paper proposes several ways for digital libraries to take account of these risks and to confront the problem of a copyrighted public domain. It also explores the responsibility of

digital libraries to provide accurate copyright information about their own products and collections, and the ways in which accuracy serves the libraries' own interests. In its analysis, the paper draws lessons from abroad.

Session 2:

Technical

Retrieval in Texts with Traditional Mongolian Script, Realizing Unicoded Traditional Mongolian Digital Library

AUTHORS Garmaabazar Khaltarkhuu, Akira Maeda, Ritsumeikan University, Japan

ABSTRACT This paper discusses our approaches to create a digital library on traditional Mongolian script using Unicode. Also we introduce system architecture of a digital library that stores books and materials of historical importance written in traditional Mongolian which contain history of 1,000 years and are important part of Mongolian culture. Specifically, we propose a technique that will allow users to search traditional Mongolian unicoded texts with keywords in modern Mongolian Cyrillic characters. To accomplish our goal, we used Greenstone digital library system and it is based on a unicoded traditional Mongolian script. We created a traditional Mongolian digital library with Golden History (Altan Tobci in Mongolian)—chronological book of ancient Mongolian Kings and their history. We approved our system's effectiveness by experiment.

Technical

Transliteration Editors for Arabic, Persian and Urdu

AUTHORS Veera Raghavendra E., International Institute of Information Technology, Hyderabad, India*;* and Lavanya Prahallad, Mostafa Fahmy, Carnegie Mellon University, United States

ABSTRACT Transliteration editors are essential for keying-in language scripts into the computer using QWERTY keyboard. Applications of transliteration editors in the context of Universal Digital Library (UDL) include entry of meta-data and dictionaries for many languages both local and International. In this paper, we propose a simple approach for building transliteration editors for International languages such as Arabic, Persian and Urdu using Unicode and by taking advantage of its rendering engine which is called Unicode rendering engine. We demonstrate the usefulness of the Unicode based approach to build transliteration editors for International Languages, and report its advantages needing little maintenance and few entries in the mapping table, and ease of adding new features to the transliteration scheme, such as adding letters. We also explain how easy it is to add any language and build a transliteration editor using Unicode and its mapping tables. We demonstrate the transliteration editor for 3 International languages and also explain how this approach can be adapted for any foreign language.

Topic?

Calligraphy Style Correlation Based on Graph Model and Its Applications

AUTHORS Lu Weiming, Zhuang Yueting, Wu Jiangqin, Zhejiang University, China (presented by Prof. Baogang Wei, Zhejiang University)

ABSTRACT As more and more works of calligraphy exists in digital library, traditional browsing and searching are not satisfying. This paper presents an algorithm for calligraphy style correlation discovery based on a graph model. We first segment the calligraphy work into characters, extract their texture features through 64 Gabor channels, and estimate the calligraphy style using a probability multi-class SVM classifier. Then we compute the style similarity between each pair of characters and select the top k neighbors to generate a data graph. Finally, we use random walk on the graph to discover the correlation among works and authors. Three experimental analyses show our proposed approach works well.

Technical

Identification and Conversion on Font-Data in Indian Languages

AUTHORS Anand Arokia Raj A., International Institute of Information Technology, Hyderabad, India; and Kishore Prahallad S., Language Technologies Institute, Carnegie Mellon University, United States

ABSTRACT To build a speech system like TTS (Text-to-Speech) or ASR (Automatic Speech Recognition) or an information extraction system like search engine, it is essential that the text has to be processed in Indian languages context. The text corpus or document database has to be in a single standard format. In this paper we discuss our effort in addressing the issues related to font-data like font encoding identification and font-data conversion in Indian languages.

Session 3:

Technical

Knowledge and Use of Digital Library Resources by Engineering Faculty Members Affiliated to Acharya Nagarjuna University, A.P. India

AUTHORS Leelavathi Navuloori, S.V. Central Library and Research Centre, India; Doraswamy M., Central Library, Siddhartha Engineering College, India

ABSTRACT An attempt has been made to determine the present status of knowledge and use of digital resources. It was observed that use of digital resources is still inadequate among the engineering faculty of the universities in the developing countries. This paper presents the findings of a survey to about the knowledge and use of digital resources by faculty members through CD-ROM databases, online databases, online journals OPAC etc available in the engineering college libraries. The subjects chosen for this study were engineering faculty members affiliated to Acharya Nagarjuna University, Andhra Pradesh, India. For evaluating study questions and data collection, the questionnaire was distributed to a random sample of 160 faculty members. The result of this survey are presented and discussed in this paper.

Technical

Analysis on Current Status of Digital Libraries in China

AUTHORS Leye Yao, Ping Zhao, Sichuan University, China

ABSTRACT On the basis of digital library (DL) definition and features, 16 Chinese library Websites, library portals and homepages were selected and investigated. The analysis and comparison focused on content set-up, digital resource type, mainstream mode, subject navigation and et al. We also investigated if the library had the Virtual Reference Service (VRS), Academic Information Resource Portal and Integrated Searching System or Platform. We found that all the selected libraries were DL or the DL initial prototypes. The electronic or digital resources were made up of two parts and the mainstream model was service-oriented. Almost all of them had their own library information portal and provided integrated searching. We also found some problems in term unification at homepage, unified retrieval platform selection between various databases, and the construction in navigation system etc.

topic??

Personalized Services in CADAL Digital Library

AUTHORS Yin Zhang, Jiangqin Wu, Yueting Zhuang, Cheng Ma, Chuan Yuan, Chunhe Wang, Zhejiang University, China (presented by Prof. Yueting Zhuang)

ABSTRACT CADAL is a great digital library project of digitizing one million books and publishing them to Internet users. It is obvious that users confront with the information overload problem when visiting the CADAL portal. Therefore, we have been concerns with providing useful and flexible personalization services to reduce the users' time and energy cost of finding interesting information. We have an extensible framework for personalization services in CADAL, including front end UI and back end module. Since the log data is plentiful and easily recorded, it is our initial step to construct the recommender system based on the massive log data. The approach implemented by us is based on two kinds of data structure: read-black header tree and prefix subtree. The results of experiments on real-world log data confirm the efficiency and excellent scalability of our approach with the large number of items and sessions.

Technical

An Overview of Techniques in the 10th Five-year CALIS Special Subject Repositories

AUTHOR Zheng-Guo Hong, Wuhan University, China

ABSTRACT This paper focuses on the related techniques in the 10th Five-year CALIS Special Subject Repositories. These special repositories adopt uniform frame of metadata sets and distributed object data, and form the mechanism of information sharing and exchange under it. The information communication is mainly implemented through the protocol of OAI, METS and CALIS-OID resolution.

Session 4:

Technical

Towards Multi-granularity Multi-facet E-Book Retrieval in China-U.S. Million Book Digital Library

AUTHORS Yonghong Tian, Tiejun Huang, Wen Gao, Peking University, China

ABSTRACT There are more than one million digitalized books (i.e. e-books) so far in China-US Million Book Digital Library Project (MBP for short). Thus an important but urgent task is to design effective and powerful tools that enable users to easily search the required information and appropriately access knowledge in the digital library. Towards this end, currently most digital libraries simply use the traditional metadata based or full-text based retrieval technologies on the e-book collection. However, there are at least two limitations of such e-book retrieval systems. (1) The granularity of retrieval results is either too big or too small, and consequently the middle granularities such as chapters, paragraphs are ignored in the traditional e-book retrieval systems. (2) The mass of retrieval results are usually ill-organized so that users often need to pay more efforts to obtain the required items. Therefore, with so many complex data in MBP, new search models and structures need to be developed that can take advantage of the particularities of e-books, access them appropriately, and provide results efficiently. To tackle this challenge, this paper introduces our multi-granularity and multi-aspect e-book retrieval approach for MBP. Firstly, a Multi-granularity Multi-facet Knowledge Network (MMKN) model is proposed to represent content from different granularities (e.g., books, chapters, pages, paragraphs and words) and different facets (e.g., time, space, etc.) to support retrieval of relevant items from a digital library collection. Then we implement a novel e-book retrieval system, called IQuery, to extract facet-related information from e-books at several granularities and then support multi-granularity e-book retrieval with more retrievable units and multi-facet navigation. Experiments were conducted to validate the efficiency and effectiveness of the proposed MMKN model, as well as the performance of IQuery. The results are encouraging, demonstrating that IQuery can provide novel and powerful capabilities for e-book retrieval in MBP.

Human Factors, Technical

Coreference Resolution using Hybrid Approach

AUTHORS Thakkar Megha Vishnuprasad, Ratna Sanyal, Indian Institute of Information Technology, Allahabad, India

ABSTRACT This work presents a novel approach to find structured information from the vast repository of unstructured text in digital libraries using Coreference Resolution. Our approach uses a salience based technique to find the antecedents of pronouns, while it uses a classifier based technique for the resolution of other noun phrases. A comparison of the proposed approach with several baselines methods shows that the suggested solution significantly outperforms them with 66.1 % precision, 58.4% recall and a balanced f-measure of 62.0%.

Policy, Technical

Model of a Digital Library in Northern Part of India; In Context to Developing Countries

AUTHOR Dr. Raj Kumar, Postgraduate Institute of Medical Education and Research, Chandigarh, India

ABSTRACT The traditional concept of a library is becoming obsolete with the emergence of the digital technology in which all the information resources are available in computer processable form and the functions of acquisition, storage, preservation, retrieval, access and display are carried out at national and international networks Medical professionals needs access to the, quick and quality health information at their desktop without the wastage of time with the help of digital library. The purpose of this study is to keep pace with new innovations of technologies, to increase effectiveness and efficiency in library services round the clock with no geographical limitations. A model of a digital library for the PGIMER to enhance the library services, quick and timely information to the medical professionals in their day to day clinical practice, medical education and research. Method and Results are based on the case study that has been undertaken in a medical institute of national importance and a medical school providing a tertiary medical care. Due to lack of proper infrastructure, inadequate finance and manpower in medical institution, the medical professionals are deprived of getting the desired information; this can lead to decline in quality health care. The present model of digital library for the PGIMER is being proposed which should includes e-resources, medical databases and electronic information products. The digital library of PGIMER could become a basis for recognizing and strengthening of library services in the region keeping in view the need and aspiration of the medical fraternity.

Policy, Technical

Language Independent Information Retrieval from Web

AUTHORS R. Seethalakshmi, Ankur Agrawal, Ranjit Ranjan, SASTRA, India

ABSTRACT Language independent information retrieval is one of the major issues in the web access by the regional population of any kind. This paper addresses the design and implementation of such information retrieval system. In this system the user is allowed to pose the query in any language and also he can retrieve the information in any other specified language. This approach encounters the design and implementation of a software_morph_parser which encompasses the natural language processing principles and retrieves the information efficiently. The software_morph_parser divides the input search text into individual words and keywords are identified. The keywords are converted into their root forms by removing all their inflexion forms and the corresponding root words are translated into the target language. The multi-lingual web database is dynamically indexed by a dyn_crawler and a search engine is invoked which searches the indexed database and ranks the pages as per the relevance to the keyword. The links are displayed to the user in the priority order of relevance. The user can click on the link and access the web page pertaining

to the required information. This system is aimed to breach the language difficulties that the regional population faces in accessing the web.

3:30 p.m. Panel on Quality Assurance

Dr. Gloriana St. Clair, Carnegie Mellon University
Dr. N. Balakrishnan, Indian Institute of Science
Dr. Jihai Zhao, Zhejiang University
Vamshi Ambati, Carnegie Mellon University
Tian Yonghong, Peking University

Sunday, November 4, Newell-Simon Hall, Carnegie Mellon University

Technology

9 a.m. Dr, Garth Gibson, Carnegie Mellon University

9:30 a.m. Lalitesh Katragadda, Google, Inc.

10 a.m. Eric Burns, Jessica Jobes, Microsoft Corporation

New Initiatives
11:30 a.m. Manohar Nadendla, MLA Govt. Andhra Pradesh

12 noon Dr. Jim Baker, institution

12:30 p.m. Indian lunch

New Initiatives
1:15 p.m. Dr. Katsu Ikeuchi, University of Tokyo

2 p.m. Parallel Sessions of Contributed Papers Locations/Moderators [topics]

Session 5:
Topic??
A Research on Establishment of Subject Database of Minguo Books
AUTHORS Yingmei Wu, Jing Huang, Songling Li, Beijing Normal University Library, China
ABSTRACT The article mainly discusses about the establishment of the subject database of Minguo books in two steps: analysis on the necessity of the establishment and the subject selections, and then emphasizes the navigation, searching and services of the database websites, based on the theory of Information Architecture. This article can be referenced for the establishment of the subject database of Minguo books.

Human Factors
Understanding the Influence of Digital Resources on Scientific Research
AUTHOR Xiaomin Liu, Library of Chinese Academy of Sciences, China
ABSTRACT Since digital library construction, it caused great change to gain information. Using the digital resources environment that Library of Chinese Academy of Sciences builds, this paper took 15 kinds of electronic periodical with higher download capacity of full text in chemistry domain as the statistical samples to analyze the download capacity of full texts during 2003-2005. It was counted how Chinese Academy of Science author quoted these 15 kinds of periodical by LCAS of Chinese Science Citation Database (CSCD) and the TSI-SCI database. It was also analyzed the relationship between downloading behavior and the citation behavior.
The selected statistics data come from 2003 to 2005. Download behavior statistics data come from full text the download capacity of 15 kinds of periodical which Publisher provided during 2002-2005. The citation statistics data come from the number of times of the CSCD origin periodical during 2003-2005 that Chinese Academy of Science scientific researcher quoted these 15 kinds of periodical in their published papers. And it was compared the domestic with oversea citation behavior through Web of the Knowledge for 2003-2005 data. Through the data statistical analysis, it may see the digital resources' influence on the scientific research behavior:

1. The periodical full text maintains high download capacity.
2. It has the similar tendency between periodical downloading and periodical citation behavior. The higher download capacity, the higher periodical number of citation times. In the same way, when download capacity is low, periodical citation quantity is also low.
3. It was significantly consistent between the CAS scientific researcher's citation behavior and the whole world scientific researcher's citation periodical behavior analyzed by JCR.
4. It was carried on the linear regression analysis, taking the periodical download capacity as the independent variable, the periodical citation

quantity as a dependent variable. It was proved that there was strong positive correlation between download behavior and citation behavior by Person regression model.

5. In the high citation first 10 papers, 80% may gain electronic version's full text directly. It possibly explains positive correlation between the high availability and the high citation behavior from the micro-layer. Digital resource is important for scientific research.

Session 6:

Technical

Building Multilingual Parallel Corpora from Scanned Pages of Digital Libraries and Gaming Techniques

AUTHORS Smita Panasa, Kranthi K., Rakesh Reddy T., MSIT, Indian Institute of Information Technology, Hyderabad, India; and Lavanya Prahallad, Carnegie Mellon University, United States

ABSTRACT There are a huge number of books that provide knowledge as well as entertainment to people. However these books cannot reach every person, as they are in languages that cannot be understood by them. It would be easier and a wonderful experience for readers if it were in their own native language. Several efforts were made to translate various books into Indian languages by a few people. However these efforts cannot be considered successful because manual process of translating books is tedious. Hence, to address this problem of translating books into various Indian languages we introduce Thingamacrib, a computer game that is enjoyable and which can be cherished always. People would love to play Thingamacrib because it is fun translating sentences. Players of Thingamacrib get entertainment, worthy gifts on playing the game and in addition to having a wonderful experience. The players not only help translating books into various Indian languages, but indirectly give joy to various others who can read their books in their native language.

Topic??

Partition Affinity Propagation for Clustering Large Scale Data in Digital Library

AUTHORS Xuqing Zhang, Fei Wu, Dingyin Xia, Yueting Zhuang, Zhejiang University, China (presented by Dr. Zhenkun Zhou)

ABSTRACT Data clustering is very useful in helping users visit the large scale of data in digital library. In this paper, we present an improved algorithm for clustering large scale of data set with dense relationship based on Affinity Propagation. First, the input data are divided in several groups and Affinity propagation is applied to them respectively. Results from the first step are grouped together in some way, and Affinity Propagation is implemented to them. Experimental results show that our algorithm, referred to as Partition Affinity Propagation, brings an encouraging effect for speeding up Affinity Propagation in clustering dense data set, while clustering accuracy are almost kept or even better.

Session 7:

Technical

Extracting Keyphrases from Books Using Language Modeling Approaches

AUTHOR Rohini U., International Institute of Information Technology, Hyderabad, India

ABSTRACT There has been a tremendous growth recently in the available digitial content. In a digital library consisting of several hundreds of thousands of books, it is clearly infeasible for a user to examine complete book to determine whether or not the document would be useful. Instead, short meta data containing descriptions of books like titles, summary, keyphrases etc., could be beneficial in helping a user in getting a quick summary about the key points in the book. We aim to perform automatic keyphrase extraction from books that can be used to get a quick overview of the key points contained. We focus on language independent approaches which can be easily be applied to other languages (other than English).

Topic??

Combining POS Taggers for Improved Accuracy to Create Telegu Annotated Tests for Information Retrieval

AUTHORS Rama Sree R. J., Kusuma Kumari P., Tirupati, India

ABSTRACT POS tagging is the process of assigning a correct POS tag (can be a noun, verb, adjective, adverb, or other lexical category marker) to each word of the sentence. POS taggers are developed by modeling the morpho-syntactactic structure of natural language text.

We attempted to improve the accuracy of existing Telugu POS taggers by using a voting algorithm. The three Telugu POS taggers viz., (1) Rule-based POS tagger, (2) Brill tagger, (3) Maximum Entropy POS taggers, are developed with an accuracy of 98.016%, 92.146%, and 87.818% respectively. An annotated corpus of 12,000 words is used to train the last two taggers.

An error analysis is made to find out the errors made by these three taggers and methods to improve the accuracy of these taggers are then examined. As a first step, a voting algorithm is proposed to build an ensemble Telugu POS tagger to get better results.

This tagged output could be used for a variety of NLP (Natural Language Processing) applications, mainly for word sense disambiguation (WSD) in retrieving Telugu documents.

Session 8:

Topic??

Patenting of Traditional Knowledge: A Study in Perspective

AUTHOR Parnika Malhotra, New Delhi, India

ABSTRACT The world today is embroiled in debate over the protection of traditional knowledge and interests of indigenous people in the third world

countries. The growing power of transnational corporations is fueled by the TRIPS in an attempt to divest the third world of its traditional knowledge. The indigenous communities are slowly waking up to the grave injustice perfected by the parasitic corporations reaping massive profits from the biological resources and traditional knowledge of the third world.

Topic??

The Implementation of Grid-Based OCR system in CADAL Project

AUTHORS Huang Chen, Xu Haiyan, Zhejiang University, China

ABSTRACT With the outburst of information amount stored in digital library, Optical Character Recognition (OCR) attracts more and more attention as the foundation of full-text search. This paper describes large-scale corporative OCR system—grid-based OCR system implemented in CADAL project, which takes full advantage of Internet and emphasizes continuously improving the text precision of a very large collection of data. The paper first analyses the process of OCR, then proposes effective solutions in the aspects of system structure, implementation steps, and data re-use, and at last discusses the problems encountered during the period of Implementation.