Presentation and Discussion:
“Digitization and Plagiarism Checking for Research Publications”
Automated Data Input and Plagiarism Checking for Research Publications

Choochart Haruechayakas, Ph.D.
Speech and Audio Technology Laboratory (SPT)
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)

Presentation outline

- Manual preparing metadata and bibliographic data
- Problem issues
- Our proposed system
- Components and functions
- Metadata and bibliographic data extraction
- Plagiarism checking
- Overview
- CopyCat Tool
**Information Life Cycle**


**Current problem issues**

- Manual preparing metadata and bibliographic data from research publications
- Labor intensive
- Time consuming
- Error prone
Current problem issues (cont'd)

- Data statistics and budget information from TCI

- Number of articles and bibl. records 38,773 / year
- Time required for data entry 10 months
- Number of staffs 30 persons
- Budget 60,000 bahts / month
  15.47 bahts / record

Our proposed system
Key advantages

- Can be customized to support
- many formats or templates
- Local languages
- Conveniently integrated with other tools
- Search engine
- Plagiarism checker

Process flow chart
Flowering time response of Nasturtium (Tropaeolum majus L.) cultivar 'Empress of India' to photoperiod, light integral and temperature using photo-biological model

Mohammed Saeed, Mohammed Rfiehil Almahboob, Abd-Al Aziz Khaleefah and towered On Dr. Ibsher

1. Introduction

Nasturtium (Tropaeolum majus L.), a wild relative of Indian Cress and Malabar Cress is native to South America and Central America. It is widely cultivated, both as an ornamental and as a medicinal plant. This research aimed at establishing flowering time

6. Conclusions

It can be concluded from the present research finding that flowering time of Nasturtium can be controlled using photoperiod and light integral in combination or alone. These factors can be manipulated to obtain desired flowering time and thereby control the flowering of this plant in various regions. The results of this study will be useful for the breeders and growers of this crop to develop varieties with desired flowering time.

Acknowledgments

The authors would like to express their gratitude to the Department of Scientific Research, Taif University, Al-Abha, Saudi Arabia for financial support.

References


Bibliographic data extraction


Authors: McDaniel, C.N., Singer, S.R. and Smith, S.M.E.
Title: Developmental states associated with the floral transition
Journal: Developmental Biology
Volume: 153
Year: 1992
Pages: 59-69

Publication format: System-friendly

• Each section has a clear section label.
• Required section labels: “Abstract”, “Keywords”, “Introduction”, “References”
• Font size and type of “Title”, “Authors” and “Affiliations” are clearly distinguished from the content.
• Authors and their affiliations are listed right after “Title”. (Not in the footer)
• Bibliographic data format follows “standard” formats.
การประชุมเพื่อสร้างเครือข่ายการพัฒนาคุณภาพวารสารวิชาการไทย ครั้งที่ 9

ศูนย์ดัชนีการอ้างอิงวารสารไทย
Thai-Journal Citation Index Centre

Publication format: System-Unfriendly

Plagiarism Checking
Plagiarism issue

- Today the Internet has become a common source for students to acquire additional knowledge outside the classroom.
- Students **copy-and-paste** the contents without doing much modification or sentence restructuring.
- The plagiarism problem has turned into a serious problem among academic and research communities.


Plagiarism in the news

**German education minister quits over PhD plagiarism**
Annette Schavan’s resignation over plagiarism ahead of election is second case to hit Merkel’s government in two years

*Staff and agencies*
*theguardian.com, Saturday 9 February 2013 13:52 GMT*

---

**Actress Kim Hye-soo Sorry for MA Plagiarism**
Actress Kim Hye-soo has apologised for plagiarising a substantial part of her master’s thesis at Sungkyunkwan University in 2001. She copied part of the thesis on actor communication verbs in from at least four books.

*Source:*
http://english.chosun.com/site/data/html_dir/2013/03/25/2013032501060.html

---

*Source:*
What is Plagiarism?

plagiarize verb ˈplā-ja-, rēz also -jē-a-

: to use the words or ideas of another person as if they were your own words or ideas

- to steal and pass off (the ideas or words of another) as one's own
- to use (another's production) without crediting the source
- to commit literary theft
- to present as new and original an idea or product derived from an existing source

Source: http://www.merriam-webster.com/dictionary/plagiarize

Which actions are considered plagiarism?

- Turning in someone else's work as your own
- Copying words or ideas from someone else without giving credit
- Failing to put a quotation in quotation marks
- Giving incorrect information about the source of a quotation
- Changing words but copying the sentence structure of a source without giving credit
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not

Source: http://www.plagiarism.org/plagiarism-101/what-is-plagiarism
How to avoid plagiarism

• Most cases of plagiarism can be avoided, by citing sources.

• Simply acknowledging that certain material has been borrowed and providing your audience with the information necessary to find that source is usually enough to prevent plagiarism.

“Proper Citation”

Source: http://www.plagiarism.org/plagiarism-101/what-is-plagiarism/

What is CopyCat?

• CopyCat is a program tool for automatically checking plagiarism in both Thai and English texts.
How does CopyCat work?

Document

Summary Report

Thai Wikipedia Custom Repository

Handling modified documents

<table>
<thead>
<tr>
<th>Text operation</th>
<th>Original document</th>
<th>Modified document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word deletion</td>
<td>A computer is a general-purpose device that can be programmed to carry out arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory.</td>
<td>A computer is a general-purpose device that can be programmed to carry out arithmetic operations automatically. A computer consists of at least one processing element, typically a central processing unit (CPU), and memory.</td>
</tr>
<tr>
<td>Word insertion</td>
<td>A computer is a general-purpose device that can be programmed to carry out arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory.</td>
<td>A computer is a general-purpose device that can be designed and programmed to carry out a set of arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory (RAM).</td>
</tr>
<tr>
<td>Word replacement</td>
<td>A computer is a general-purpose device that can be programmed to carry out arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory.</td>
<td>A computer is a general-purpose device that can be programmed to perform arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, normally a central processing unit (CPU), and some type of memory.</td>
</tr>
<tr>
<td>Sentence alternation</td>
<td>A computer is a general-purpose device that can be programmed to carry out arithmetic or logical operations automatically. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory.</td>
<td>Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory. A computer is a general-purpose device that can be programmed to carry out arithmetic or logical operations automatically.</td>
</tr>
</tbody>
</table>

Source: https://en.wikipedia.org/wiki/Computer
CopyCat's Workflow

1. Upload file
2. Analysis
3. Report

Thank You
ประวัติโดยย่อ

ชื่อสกุล: ดร. ชูชาติ หฤไชยสกุล
สถานที่ทำงาน: ห้องปฏิบัติการวิจัยเทคโนโลยีเสียง หน่วยวิจัยวิทยาการสารสนเทศ ศูนย์เทคโนโลยีสารสนเทศและคอมพิวเตอร์แห่งชาติ สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ อุทยานวิทยาศาสตร์ประเทศไทย ถนนพหลโยธิน ตำบลบึงช้าง อำเภอคลองหลวง จังหวัดปทุมธานี 12120
โทรศัพท์: 0-2564-6900 ต่อ 2240
อีเมล: choochart.haruechaiyasak@nectec.or.th

CAREER OBJECTIVES
- Perform R&D in Natural Language Processing, Information Retrieval, and Data / Web / Text Mining
- Support IT industry by providing technological know-hows in the fields related to search engine and data analytics with an emphasis on business, marketing, e-commerce, and Internet applications.
- Disseminate knowledge and experience to academic institutions.

EDUCATION
June 2003 University of Miami, Miami FL, U.S.A.
   Doctor of Philosophy in Electrical and Computer Engineering
December 1997 University of Southern California, Los Angeles CA, U.S.A.
   Master of Science in Computer Engineering
May 1995 University of Rochester, Rochester NY, U.S.A.
   Bachelor of Science in Electrical Engineering (with Distinction)
   Certificate of Management Studies in Accounting and Finance

HONORS and AWARDS
1990 – 2003 Scholarship from the Royal Thai Government
December 2002 Jury’s Best Graduate Student Award, University of Miami
May 2003 Outstanding Graduate Student from College of Engineering, University of Miami

WORK EXPERIENCE
2014-2015 Project manager on Talent Mobility
2013-2014 Project manager on Expert Finder System for TNRR
2013-2014 Project manager on Social Web Content Mining at NECTEC
2009-2011 Project manager on Thai Web Archive and Text Mining at NECTEC
2007-2009 Project manager on ThaiReSearch at NECTEC
2004-2010 Lecturer for course: Decision Support Systems (DSS)
Faculty of Information Technology, King Mongkut’s University of Technology North Bangkok
2007 - present  Lecturer for course ECS 218 Data Structures and Algorithms
Sirindhorn International Institute of Technology (SIIT)

PATENTS
1. A Portable Electronic Apparatus with Thai Word Prediction Function (pending for approval)
2. Automatic Online Help Desk System (pending for approval)
3. A Method and System for Opinion Mining (pending for approval)
4. A Text Mining System for News Article Corpus (pending for approval)
5. A Text-Based Indexing Method for Image Retrieval with Color and Tone (pending for approval)
6. An Information Filtering System for Thai Language (Petit Patent No. 8105)
7. A Portable Electronic Apparatus with an abbreviated typing function (pending for approval)
8. An Apparatus for Displaying an Adaptive Emoticon According to the Emotion in Texts (pending for approval)
9. An apparatus for Thai Word Segmentation and Normalization (pending for approval)

SYSTEM PROTOTYPES
5. ABDUL (Artificial Buddy U Love), Online Personal Agent, available at: http://www.abdul.in.th/demo/
10. Thai Unknown Word Corpus Construction
12. Thai Q-Cor: Thai Query Correction, source available at: http://www.sansarn.com/thaiquery/

RESEARCH INTEREST
- Natural Language Processing (NLP)
- Data / web / text mining
- Opinion mining and sentiment analysis
- Information Retrieval (IR) and search engine
- Information filtering and recommender systems
- Data mining applications in e-commerce, web intelligence, management, finance and marketing
- Big data analytics and visualization
PROFESSIONAL SERVICES

2008-2014  Program Committee: IEEE Int. Conf. on Multimedia (ISM)
2009-2014  Program Committee: The Int. Joint Conf. on Computer Science and Software Engineering (JCSSE)
2007-2014  Program Committee: The National Conf. on Computing and Information Technology (NCCIT)
2011-2014  Program Committee: The Int. Conf. on Computing and Information Technology (IC2IT)
2005-2014  Program Committee: IEEE Int. Conf. on Information Reuse and Integration (IRI)
2008-2014  Program Committee: ECTI-CON
2009-2014  Program Committee: The Int. Conf. on Advances in Social Network Analysis and Mining (ASONAM)
2007-2014  Program Committee: IEEE Int. Workshop on Multimedia Information Processing and Retrieval (MIPR)
2008-2014  Program Committee: The Int. Conf. On Multimedia and Ubiquitous Engineering (MUE)
2009-2014  Program Committee: The Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

SELECTED PUBLICATIONS