

Next Generation Digital Libraries: Mining Our Own Scholarly Documents

KAN Min-Yen

Web IR / NLP Group (WING)



NUS
National University
of Singapore

School of
Computing

29 May 2014

4th SEA V

Varun Sivamani, a FYP graduate of WING, launched *Semantics3* with three friends, web. Since then the company has gone to attract paying customers and has recent startup to be selected for incubation with YCombinator. Congratulations, *Semantics3*



Home

News

People

Downloads

Projects

Publications

Meetings

Web Se

Research Topics in WING

Digital Libraries << Focus Today

Natural Language Processing

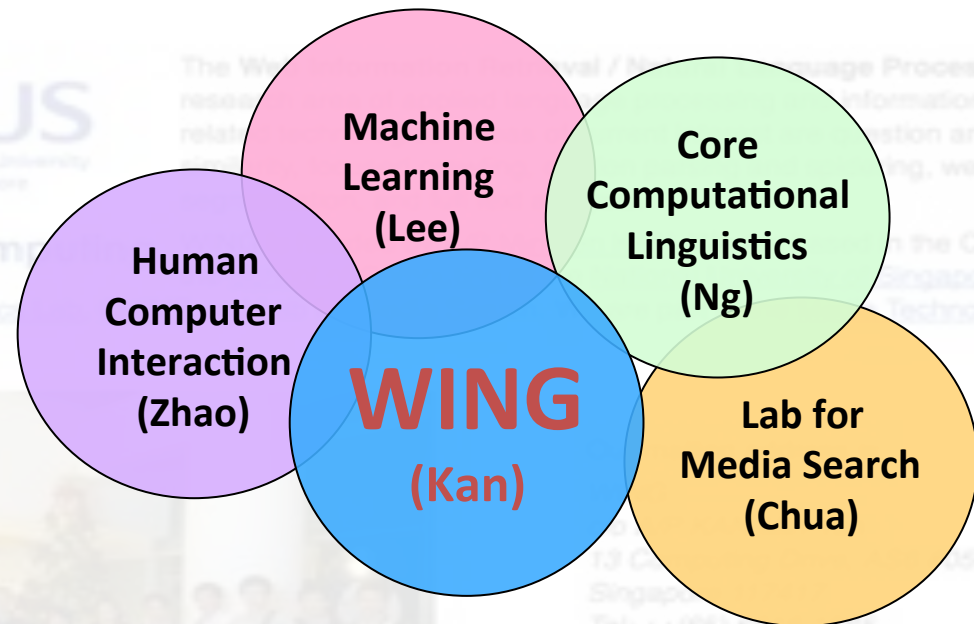
- Machine Translation
- Information Extraction
- Automatic Summarization

Information Retrieval

- Understanding Users
- Multimedia Alignment

Human Computer Interaction

- Design of Browsing / Searching User Interfaces



Password

Sushi (Rochester Mall). Click on the image to see others.

We read papers, lots of papers!

How do we make sense of this knowledge?

By just printing them out proceedings?

Scholarly digital libraries, circa 2014

Characteristics:

- PDFs and metadata of papers
- Searchable

The screenshot displays the PubMed website interface. At the top, there is a search bar with the text "Search PubMed" and a "Go" button. Below the search bar, there are several navigation tabs: "Limits", "Translations", "History", "Clipboard", and "Details". A prominent green box in the center contains a notice about the NIH Public Access Policy, asking "Does NIH fund your work?" and providing instructions on how to make manuscripts available in PubMed Central. The notice includes a link to "pubmedcentral.nih.gov". To the left of the main content, there is a sidebar with various links and categories, including "About Entrez", "Entrez PubMed", "PubMed Services", and "Related Resources". At the bottom of the page, there is a footer with the text "Works in the Open Access (OAJ) & Health (OAH) Department of Health & Human Services (DHHS) Division of Information & Statistics".

Case in point: ACM Portal

Characteristics:

- Also has author statistics, co-authorship list and citation records
- Emphasize impact of author, vanity search

ACM DIGITAL LIBRARY

Dipri Prasad Mukherjee
Authors: [Add personal information](#)

Affiliation history
 • [Indian Statistical Institute, Kolkata](#)
 • [University of Alberta](#)

Bibliometrics: publication history

| | |
|-------------------------------|-----------|
| Publication years | 1992-2013 |
| Publication count | 32 |
| Citation Count | 79 |
| Available for download | 5 |
| Downloads (6 Weeks) | 19 |
| Downloads (12 Months) | 125 |
| Downloads (cumulative) | 453 |
| Average downloads per article | 90.60 |
| Average citations per article | 2.47 |

SEARCH: 32 search results
10 per page | Sort by: year

ROLE: Author only

AUTHOR'S COLLEAGUES: See all colleagues of this author

SUBJECT AREAS: See all subject areas

KEYWORDS: See all author supplied keywords

FEEDBACK

AUTHOR PROFILE PAGES: Project background, Authorizer Service

BOOKMARK & SHARE

2013

1 [A design-of-experiment based statistical technique for detection of key-frames](#)
 Snehasis Mukherjee, Dipri Prasad Mukherjee
 February 2013 **Multimedia Tools and Applications**, Volume 62 Issue 3
 Publisher: Kluwer Academic Publishers

Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Citation Count: 0

In this paper decision variables for the key-frame detection problem in a video are evaluated using statistical tools derived from the theory of design of experiments. The pixel-by-pixel intensity difference of consecutive video frames is used as the ...

Keywords: Design of experiment, Gestalt, Helmholtz principle, Key-frame, Meaningfulness, Video summarization

2012

2 [Synthesis of emotional expressions specific to facial structure](#)
 Swapna Agarwal, Maitreya Chatterjee, Dipri Prasad mukherjee
 December 2012 **ICVGIP '12: Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing**

Publisher: ACM
 Full text available: PDF (5.40 MB)

Bibliometrics: Downloads (6 Weeks): 7, Downloads (12 Months): 41, Citation Count: 0

Imposing expressions on expression-neutral human face images is an interesting application of human-computer-interaction, animation, entertainment and other such fields. The objective of this paper is to impose one of the six prototypic emotional expressions ...

Keywords: PCA, expression intensity, facial expression synthesis, facial structure, nonlinear mapping

Rexa

- From the machine learning community
- Papers, authors and grants as first-order objects
- Requires login

The screenshot shows the Rexa.info website interface. At the top, there is a navigation bar with links for 'Research', 'People', and 'Connections'. A search bar contains the text 'multiword'. Below the search bar, there are several search results listed:

- 1. Multiword Expressions: A Pain in the Neck for NLP**
Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Daniel Flickinger
CICLing, 2002
Multiword expressions are a key problem for the development of large-scale, linguistically sound natural language processing technology. This paper surveys the problem and some currently available analytic techniques. The various kinds of multiword expressions should be analyzed in distinct ways, including listing 'words with spaces', hierarchically organized lexicons, restricted combinatoric ... comprehensive analysis of multiword expressions must employ both symbolic (23 citations)
- 2. New Cryptographic Primitives Based on Multiword T-Functions**
Alexander Kirov, Ad Shamir
FSC, 2004
A T-function is a mapping from n-bit words to n-bit words in which for each $0 \leq i < n$ (bit) of the output can depend only on bits $0, 1, \dots, i$ of the input. All the boolean operations and most of the numeric operations in modern processors are T-functions, and their compositions are also T-functions. In earlier papers we considered 'crazy' T-functions such as ... (4 citations)
- 3. Multiword expressions: linguistic precision and reusability**
Ann Copestake, Fabre Lambeau, Aine Vallencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, Daniel Flickinger
University of Cambridge Computer Laboratory
This paper discusses the approach to multiword expressions being adopted in the LinGO English Resource Grammar (<http://lengo.stanford.edu>), a broad-scale bidirectional grammar of English in the HPSG framework. We discuss how the lexicon of multiword expressions is encoded in a database and describe the implications for building a reusable lexical resource (5 citations)
- 4. Extracting Multiword Expressions with A Semantic Tagger**
Scott S. L. Phao, Paul Reyson, Dawn Archer, Andrew Wilson, Tony McEnery
Automatic extraction of multiword expressions (MWE) presents a tough challenge for the NLP community and corpus linguists. Although various statistically driven or knowledge-based approaches have been proposed and tested, efficient MWE extraction still remains an unsolved issue. In this paper, we present our research work in which we tested approaching the MWE issue using a semantic field annotator. We use an English semantic tagger (UGAS) developed at Lancaster University (2 citations)
- 5. Multiword unit hybrid extraction**
Ghan V. Das
In Workshop on Multiword Expressions of the 41st ACL meeting, Sapporo, Japan, 2003 (1 citation)

Rexa's Service

- Embed a layer serviced by Rexa
- Like Trackback for blogs

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Annealing Data Set
Download: [Data Folder](#), [Data Set Description](#)

Abstract: Steel annealing data

| | | | | | |
|----------------------------------|--------------|-----------------------------|-----|--------------|----------|
| Data Set Characteristics: | Multivariate | Number of Instances: | 700 | Area: | Physical |
|----------------------------------|--------------|-----------------------------|-----|--------------|----------|

Papers That Cite This Data Set¹:

Rexa.info
Research • People • Contributions

Qingping Tao Ph. D. [MAKING EFFICIENT LEARNING ALGORITHMS WITH EXPONENTIALLY MANY FEATURES](#)
Qingping Tao A DISSERTATION Faculty of The Graduate College University of Nebraska In Partial Fulfillment of Requirements. 2004. [\[View Context\]](#)

Yuan Jiang and Zhi-Hua Zhou. [Editing Training Data for kNN Classifiers with Neural Network Ensemble](#). ISNN (1). 2004. [\[View Context\]](#)

Jihoon Yang and Rajesh Parekh and Vasant Honavar. [DistAI: An inter-pattern distance-based constructive learning algorithm](#). *Intell. Data Anal.* 3. 1999. [\[View Context\]](#)

Pedro Domingos. [Knowledge Discovery Via Multiple Models](#). *Intell. Data Anal.* 2. 1998. [\[View Context\]](#)

Zhi-Hua Zhou and Xu-Ying Liu. [Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem](#). [\[View Context\]](#)

James J. Liu and James Tin and Yau Kwok. [An Extended Genetic Rule Induction Algorithm](#). Department of Computer Science Wuhan University. [\[View Context\]](#)

Revisiting the characteristics

- Led by computer science
- Digital library as a monolithic server
- Automation and scalability
- Document access, not document use
 - Focus on rights management → OpenURL
 - Focus on identifying records → DOI

Centered on the authors!
What happened to the readers?

They got ...
...left out?

We still end up reading paper
Or digital copies

We treat papers as objects...
... not as a organized collection
of ideas

CiteULike

- Web 2.0 style – powered by users
- Upload your own metadata
- Got so big, publishers allow ingestion directly



- Firefox management of reference data
- Scrapes info from specific webpages
- Loads manager in place
- EndNote, other systems also offering such functionality

The image shows the Zotero website and a Firefox browser window. The website header includes the Zotero logo, navigation links (Home, Support, Forums, Developers, Blog, About), and a prominent red 'Download' button. Below the download button, it states 'Latest version: 1.0.7' and provides instructions for installation and minimum requirements (Firefox 2.0+ or 3.0, Netscape Navigator 9.0, or Flock 0.9.1). A 'ZOTERO 1.5 SYNC PREVIEW' section is also visible.

The browser window shows the Zotero extension interface. A red arrow points from the 'Download' button on the website to the Zotero interface in the browser. An orange box highlights the Zotero interface in the browser window. The interface shows a search for 'Shakespeare: the invention of the human' and a list of results. A 'Saving item...' dialog box is open, showing the item details and a 'Save' button.

FEATURES

- Automatic capture of citation information from web pages
- Storage of PDFs, files, images, links, and whole web pages
- Flexible notetaking with autocave
- Fast, as-you-type search through your materials
- Playlist-like library organization, including saved searches (smart collections) and tags
- Platform for new forms of digital research that can be extended with other web tools and services

ZOTERO NEWS

- Become A Zotero Trainer novembre 11, 2008
- Final Sync Preview Release Zotero's Notes Get Rich or Die Tryin' octobre 31, 2008
- Official Statement octobre 29, 2008
- More Upcoming Features: Browse Your Zotero Library Online septembre 30, 2008

Merging the two

- Librarians and end users have thought about the readers
 - Supporting re-use (Citation management)
 - Sharing metadata through group management
- Computer scientists, about the author
 - Prestige, finding authors and papers

But still not about **using** the document!

- Beginning researchers: undergraduate students
- Professors: managing students

Next Gen DL Elements

1. Summarizing Documents
2. Aligning Papers to their Presentations
3. Rediscovering Document Structure

Summarizing Scientific Documents

Reading and understanding is hard.

Scientific discoveries are written for peer experts to understand, but not for novices to learn.

Idea: Use NLP and AI to digest scientific articles to make them easier to understand for beginning students

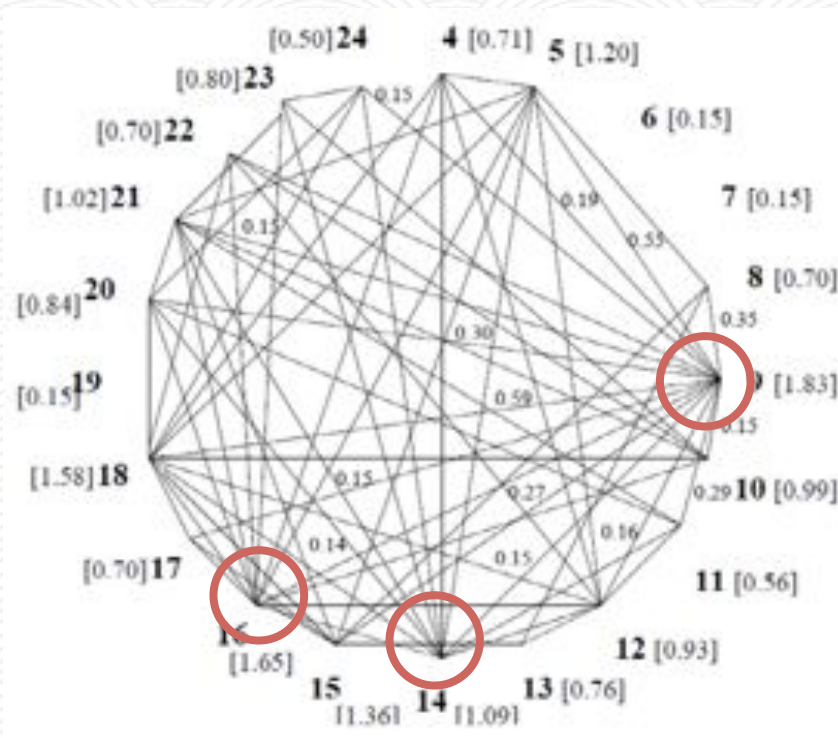
Document Summarization

- First try: To build a summary of a document
- To extract k sentences from a document of size n
- To rank the n sentences from a document with respect to some criterion

- Machine learned document summarization: Find features that help to select good sentences for a summary. Train from summaries.

Finding central sentences: TextRank

Build a graph of how much word “overlap” is there between sentences in a text.



Realize that the sentences that are more “central” are probably more important.

What about *scientific* documents?

Recognizing Human Action at a Distance in Video by Key Poses

Sachin Mukherjee, Sojoy Kumar Bhowm, Member IEEE, and Dipri Prasad Mukherjee, Senior Member IEEE

Abstract—In this paper we propose a graph theoretic technique for recognizing human actions at a distance in a video by modeling the visual scenes associated with poses. The proposed methodology follows a bag-of-words approach that starts with a large vocabulary of pose related words and derives a robust and compact codeword of key poses using centrality measures of graph connectivity. We introduce a “meaningful” threshold on centrality measures that selects key poses for each action type. Our contribution includes a novel pose descriptor based on Histogram of Oriented Optical Flow (HOOF) evaluated in a hierarchical fashion on a video frame. This pose descriptor combines both pose information and motion patterns of the human performer into a multidimensional feature vector. We evaluate our methodology on four standard activity-recognition datasets demonstrating the superiority of our method over the state-of-the-art.

I. INTRODUCTION

Recognizing action of a distant human performer in video is useful in many applications ranging from wide-area video surveillance to evaluating game tactics by autonomous vision systems. The challenge primarily lies in extracting distinct visual patterns from the gestures of human performer given the fact that they look like appearance of the distant human figure [1]–[3]. If the height of the human does not have much room for modeling the limbs separately, the only reliable cue in such case is the pose specific information and we look on the motion patterns of the poses to derive pose descriptors. Our emphasis on poses is based on the premise that human actions are composed of repetitive motion patterns and a sparse set of key poses often suffice to characterize an action. The proposed methodology follows the bag-of-words approach [4], [5]. Suppose we have N documents containing words from a vocabulary of size M and an $M \times N$ co-occurrence table T is formed, where $T(i, c, d)$ counts the number of occurrences of a word i , in document d . In our approach, “word” refers to a vocabulary of human poses obtained from pose descriptor (codeword) of each frame in a video is represented by a vector called pose descriptor of the human figure and a “document” corresponds to the entire video sequence of a particular action type (i.e., running, walking, jumping, etc.). We build a particular action descriptor by constructing a histogram of pose “words” occurring all through the video. Before describing the motivation, we first present the related works.

Sachin Mukherjee, Sojoy Kumar Bhowm and Dipri Prasad Mukherjee are with the Electronics and Communication Sciences Unit, Indian Institute of Technology, Kharagpur, India. E-mail: mukherjee@ece.iitkgp.ac.in, sbk@iitkgp.ac.in.

A. Related works

The literature in the field of human action recognition in image or video usually have two broad classifications either they focus on low and mid-level feature extraction (i.e., template based approaches) or they model the high level information among the features (i.e., model based approaches). Template based approaches train variable classifiers, or are compared to a set of event templates for recognition [6], [1]. Tefli et al. [6] have proposed a diagonal feature based learning framework where mid level shape features are constructed from low level gradient features using AdaBoost algorithm. In a typical model based approach, Mori et al. have proposed a learned probabilistic model to represent human body parts in an image, where the action is recognized by matching the state pointers in the image with the target action [6], [6]. Similarly, Chung et al. have used the silhouette of the body parts to represent the shape of the performer [7]. Recently, the bag-of-words model is being used to recognize actions in videos [8], [9].

Shih et al. have used a vocabulary of local spatio-temporal features (called cuboids) and a vocabulary of spine images (to capture the shape deformation of the actor) by considering actions as 3D objects [8]. Nibbelk et al. also use some space-time interest points on the video as features (visual words) [1]. The algorithm of Sathya et al., automatically learns the probability distributions of the visual words using graphical models like probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). Meng et al. [9] have proposed a generative mixture model for video sequences using velocity history of tracked key points, but they specifically focus on the high resolution video. As opposed to “collection of words” representing each frame [1], Mori et al. [7] have represented each frame as a single “word” achieving state-of-the-art accuracy in low resolution videos. There are some good efforts of combining global and local features for action recognition [6], [10], [11], [12]. Han et al. have introduced a multiple kernel based classifier to automatically select and weigh both low level and high level features for action recognition where body part detectors are used to “weight” in their bag-of-words model [6]. Wang et al. have combined global and local features and applied Hidden Conditional Random Field (HCRF) [10], [12] model for action recognition. Later Wang et al. have learned the parameters in HCRF in a max-margin framework [11], [12] and called the new model as Max-Margin Hidden Conditional Random Field (MHCRF). Liu et al. have proposed a methodology that does not deal with the temporal behavior of the words and

Let’s look at an example from Prof. Mukherjee

I. They have specific presentation conventions

I. INTRODUCTION

A Related Works

B Motivation and Overview

...

Abstracts!

Aim

Abstract—In this paper we propose a graph theoretic technique for recognizing human actions at a distance in a video by modeling the visual senses associated with poses. The proposed methodology follows a bag-of-word approach that starts with a large vocabulary of poses (visual words) and derives a refined and compact codebook of key poses using centrality measure of graph connectivity. We introduce a ‘meaningful’ threshold on centrality measure that selects key poses for each action type. Our contribution includes a novel pose descriptor based on Histogram of Oriented Optical Flow (HOOF) evaluated in a hierarchical fashion on a video frame. This pose descriptor combines both pose information and motion pattern of the human performer into a multidimensional feature vector. We evaluate our methodology on four standard activity-recognition datasets demonstrating the superiority of our method over the state-of-the-art.

Methods

Evaluation

These are the “ground truth” for a summary of a single paper

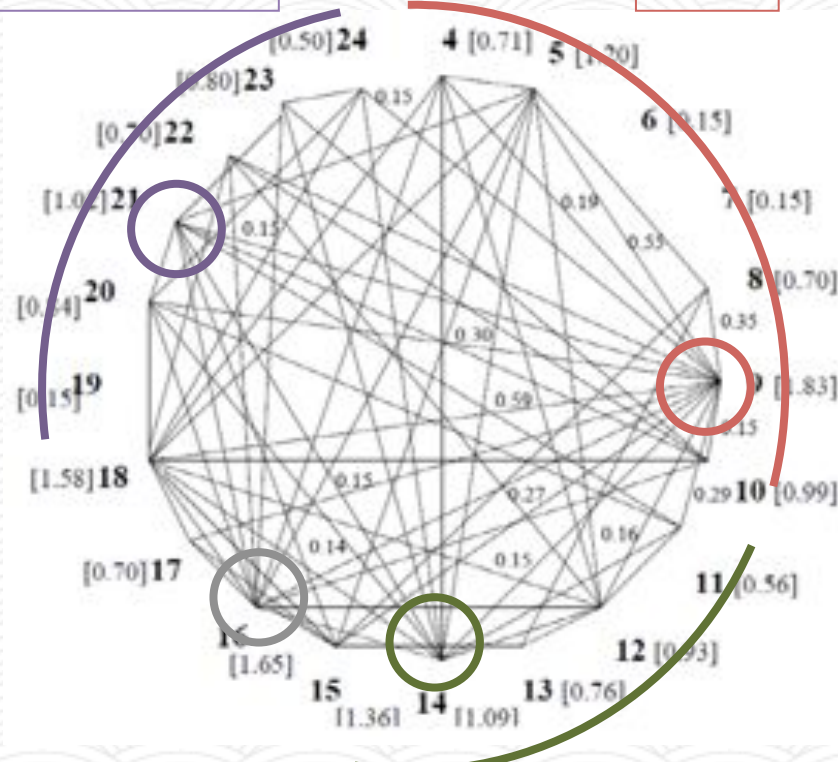
Want to select sentences that overlap significantly with the abstract

But even here there is structure within an abstract

Revisiting TextRank

Evaluation

Aim

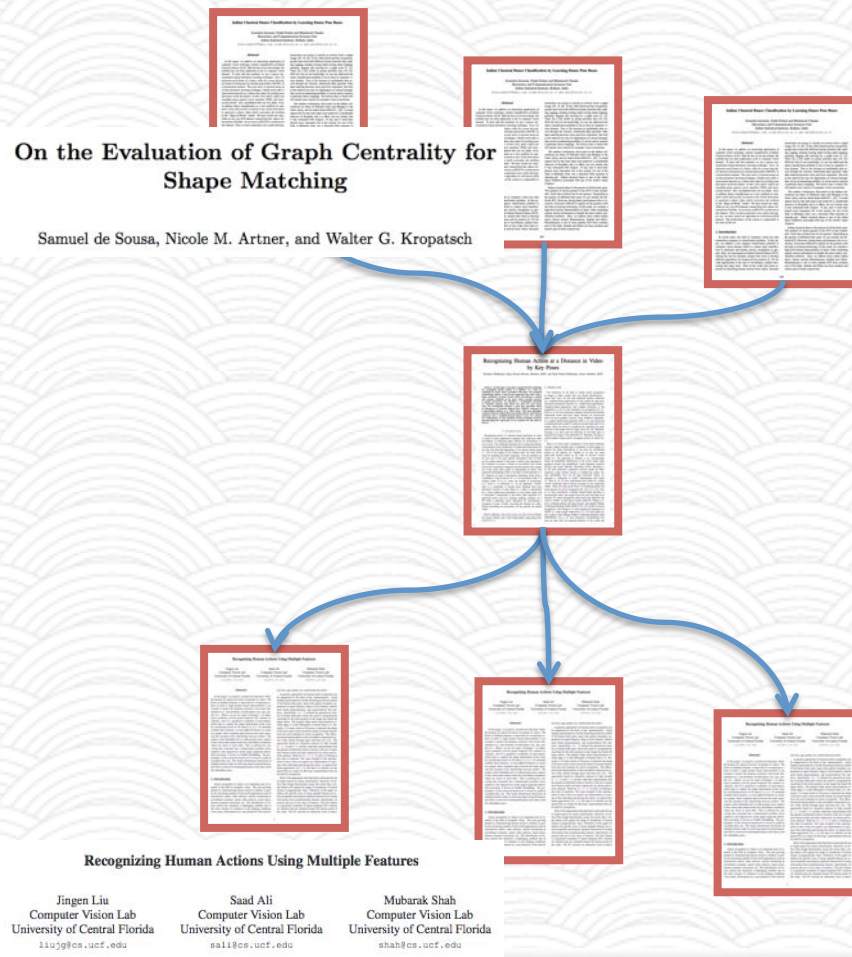


Methods

Capitalize on the conventional structure in documents.

- Identify the logical structure of a scientific document
- Find the best sentences within the sections of the document

What about *scientific* documents?



- They have specific presentation conventions
- They have references and citations

Citation Sentences



On the Evaluation of Graph Centrality for Shape Matching

Samuel de Sousa, Nicole M. Artner, and Walter G. Kropatsch



Mukherjee et al. [13] present an application of centrality in human action recognition. They employ centrality to create a compact codebook out of a large vocabulary of poses (bag-of-word approach). Cukierski et al. [7] use centrality

- Often describe a paper from the community's point of view
- A representation of a key point of a work

A star and his fans



Photo Credits: [Bollywood Glamour](#)

Citation sentences and in-article sentences have complementary purposes

- Results and evaluations usually not mentioned in citation sentences.
- Sentences in a paper that describe its method are usually too detailed for a summary.

Ambiguous Poses

Aligning Papers to their Presentations

Ambiguous Poses



We attend conferences in part to help learn from each other.

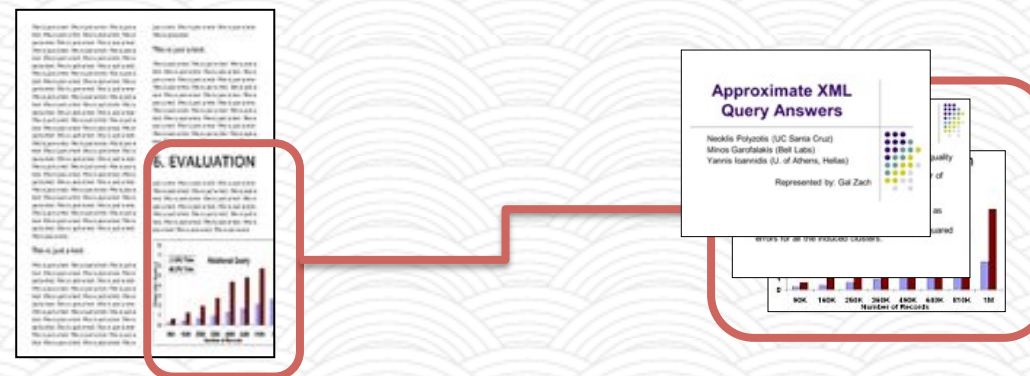
A key artifact is the slide presentation, which often summarizes the work in an accessible manner.

But they:

- Are not detailed enough
- Miss important technical details

Idea: Use both together

Better to juxtapose both media together in a fine-grained manner.



Output: an alignment map

A prototype

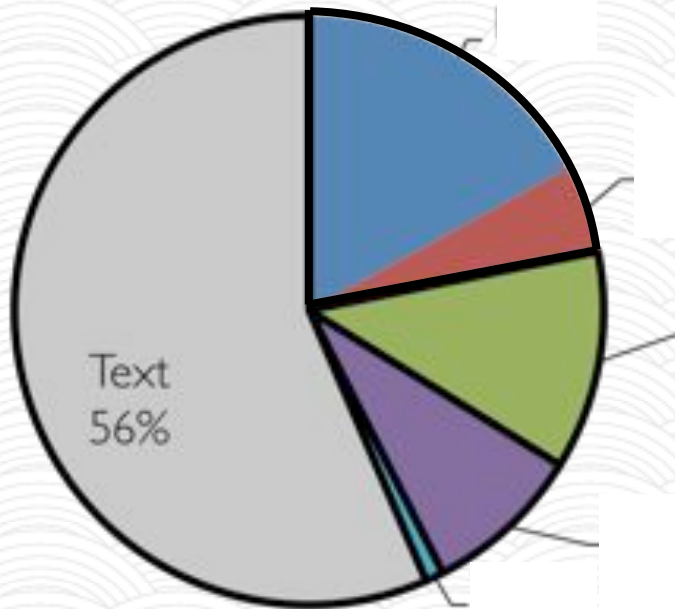
The image displays two side-by-side screenshots of a Beamer presentation. Both slides have a blue header with the title 'Approximate XML Query Answers' and authors 'Neoklis Polyzotis, Minos Garofalakis, Yannis Ioannidis'. The left screenshot is slide 7 of 9, titled '6. EXPERIMENTAL STUDY'. It features a table of contents on the left and text describing an experimental study of T REE S KETCH. The right screenshot is slide 39 of 53, titled 'Approximate Answers'. It contains a line graph for 'IMDB (~102K Elements) Avg. Result Size 3,477 tuples' comparing 'T REE S KETCH' and 'TWIG-XS KETCH'. A vertical sidebar on the right of the second slide contains text: 'the effects of our novel T REE S KETCH synopsis as a practical solution for general approximate answer to complex twig queries'.

Document in focus

Slides in Focus

Demographics of an Existing Dataset

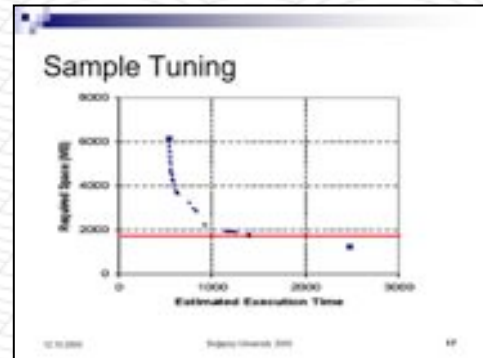
Prevalence of Slide Types



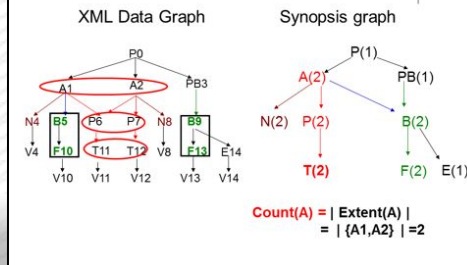
Thank You!

Agenda

- Problem
- Solution: Conceptual Partitioning Monitoring (CPM)
- Extensions of the Solution
- Performance Analysis
- Conclusion



Synopsis Model - Example



Combining Evidence

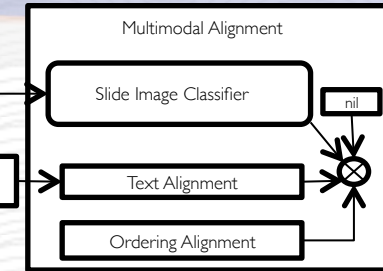
Represent each of the three sources as a probability distribution or preference

1. Text Similarity
2. Linear Ordering
3. Visual Content

Handle obvious exceptions.

Weight distributions together to find most likely point as alignment.

System Architecture

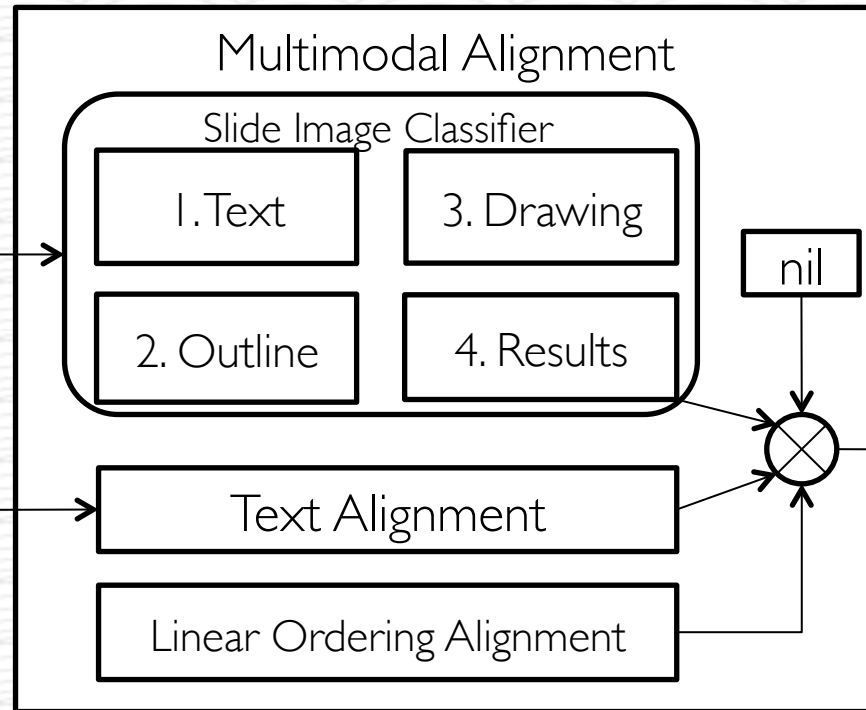


Input: Presentation

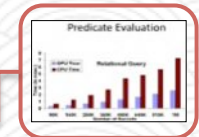


Input: Document

Pre-processing

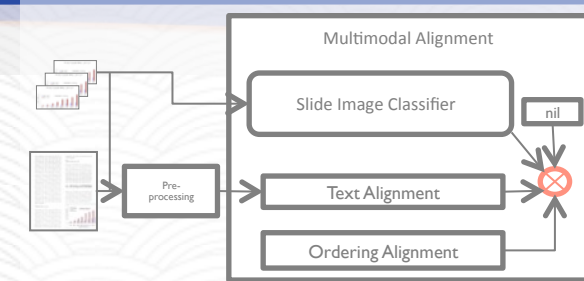


Output: Alignment map



Current architecture. Slightly different from published paper.

Multimodal Fusion



- Input for each slide:
 1. Text Alignment Vector $\rightarrow V_{Ts}$
 2. Ordering Alignment Vector $\rightarrow V_{Os}$
 3. Class assigned from image classifier \leftarrow Uses Histogram of Oriented Gradients (HOG)
- Define 3 weights as: $W_{Ts} + W_{Os} + W_{nil} = 1.00$
- Tune weights according to image classes
- Apply *Nil* classifier
- Output for each slide: Final Alignment Vector $\rightarrow FAV_s$

Results



A close-up photograph of a bull's head, focusing on its large, curved horns and the textured, reddish-brown skin of its face. The background shows a green field, a white fence, and a red barn. A semi-transparent white box with rounded corners is overlaid on the image, containing the title text.

Rediscovering Document Structure

Publishers want authors to annotate the structure of the document in markup.

But history has shown time and time again that people don't always follow instructions.

We need tools that infer the logical document structure from finished documents

“Creating the cow from the hamburger” – Steve Pfeiffer

ParsCit

putting papers back together since 2008

- Reference String Parsing and Logical Document Structure Discovery Tool
- Other subsystems built on top:
 - Technical terminology finder,
 - Author and affiliation matcher, and
 - Document summarization.

+ Toggle Table of Contents

[Back to the INWIG home page](#) | [Back to ParsCit/Parser web services](#)
[Download](#) | [Web Service](#) | [Web-based Demonstration](#) | [Publications](#) | [Get Standard Input and Sample Output](#) | [Drive Members](#) | [Contact Us](#)



ParsCit: An open-source CRF Reference String Parsing Package

This is the home page of the ParsCit project, which performs reference string parsing. It is architected as a supervised machine learning procedure that uses Conditional Random Fields as its learning mechanism. You can download the code below, parse strings online, or send batch jobs to our web service (coming soon). The code contains both the training data, feature generator and shell scripts to connect the system to a web service (used here too).

Some definitions (thanks to Robert Dale)

Reference String:
A text string in the bibliography or reference section of a work, usually at the end of the document that refers to a unique document. Usually occurs with other reference strings that point to other documents. May also appear as footnotes.

Citation:
A text string (usually explicit) in the document body that points to a corresponding reference string at the end of the document. Several citations may co-refer to a single reference string.

This project deals with the problem of parsing the reference strings. Other projects related to ParsCit (some here in INWIG, some elsewhere) deal with parsing the headers of the document (i.e., information on the title page) and with identifying and linking citations to reference strings.

Download

You can download the open source code for ParsCit here (coming soon). The source requires you to re-compile the CRFPP source code and assumes that perl is installed on your system and can be invoked using `perl` (must be in your path).

- Current version: [0.00017](#) minor changes (improved models and multilingual support), see [Changelog](#)
- Other versions: [0.00002](#) first public release. Comes with precompiled linux binaries for CRF++; Beta release [0.00010](#)
- [CRF++](#): A conditional random fields toolkit that you may need to install, if the compiled one does not work for you. As newer versions store the learned model in different file formats we recommend that you use version 0.48.

Web Service

More NLP services are now being made available on the web. Following this trend you can send your plain text citations to use via our web service. We will parse these for you free of charge (as and when time and processing power allows, these processes are done with lower priority).

N.B. We keep logs of what's parsed in these demos, to improve the accuracy and productivity of ParsCit. If you'd like these to be kept private or you find you use this service a lot, why not install a local copy of ParsCit for yourself? If you do, please let us know where you are so we acknowledge you.

This repository - Search or type a command - Explore Gist Blog Help knmnyn + X

knmnyn / ParsCit Unwatch 12 Unstar 40 Fork 10

An open-source GRF Reference String Parsing Package <http://wing.comp.nus.edu.sg/parsCit> — Edit

213 commits 2 branches 0 releases

branch: master ParsCit / +

Update to the Troubleshooting section on the webpage

wing.nus authored on Oct 23, 2013

| File | Commit Message | Time Ago |
|----------------|--|--------------|
| bin | Issue 13. The word 'Note(s)' is being misinterpreted as a section hea... | 7 months ago |
| crfp | Revert "Issue 13. The word 'Note(s)' is being misinterpreted as a ser... | 7 months ago |
| demodata | Revert "Issue 13. The word 'Note(s)' is being misinterpreted as a s... | |
| doc | Update to the Troubleshooting section on the webpage | |
| lib | fix conflicts | |
| resources | Fix CR LF line endings | |
| test | New ParsCit version: 110505 | |
| wsdl | Thing: commit before sync | |
| .gitignore | incorporate BibloScript | |
| CHANGELOG | Small change: CHANGELOG | |
| COPYING | Initial setup | 4 years ago |
| COPYING.LESSER | Initial setup | 4 years ago |
| INSTALL | Update INSTALL | 6 year ago |
| README | Changed documentation | 3 years ago |

Code Issues Pull Requests

Success stories:

- Used in CiteSeerX – over 20M citations parsed as of 2008
- Used by Mendeley in their software libraries
- Used by over 20 documented DL projects

Clone in Desktop Download ZIP

It's open source (LGPL)!
Go try it out!

Search for "ParsCit" on the web.

Conclusions

- Building open-source tools for authors and readers
- Studying users as they interact with the DL
 - Literature Review – [Single Document Summarization](#), Survey paper generation
 - Downstream Data Mining – Author affiliations, Technical terms, Zoning of the paper, Key metadata, [Scholarly document structure](#)



WING and visitors, Dec 2013

My question to all of you: What will the “publishing” model look like in 5-10 years?

THANK YOU