How Much Does $1.7 Billion Buy?

A Comparison of Scientific Journal Articles to Their Pre-print Versions
Global Trends in Scientific Output

- Global STM publishing market $25.2 billion
  - 55% of STM comes from USA
  - 40% of STM is for journals ($10 billion)
    - 68-75% is coming out of libraries’ budgets

USA is the largest contributor in terms of global output of research papers: 23%
How Much Does $1.7 Billion Buy?

@farbthink, @liber8er, @peterbroadwell, @mart1nk1e1n

#CNI15F, Washington, DC, 12/15/2015

UC Publication Impact

Across all disciplines, the UC FWCI average is 2.15; the U.S. FWCI average is 1.48.
What is the $ of Knowledge?

Prices Set by Profit Maximizing Publishers are Determined NOT by costs, but by what the market will bear.

Pre-print v. Final Published?

**arXiv**
- The annual budget for arXiv is circa $826,000 for 2013 - 2017

**Final Published**
- English language STM journals: $10 billion in 2013
Role of Publisher*

- Entrepreneur
- Copyediting
- Tagging
- Marketer
- Distributor
- E-Host

*STM Report 2015: An Overview of Scientific and Scholarly Publishing
Michael Ware and Michael Mabe
Working Assumptions

• If the publishers’ argument is valid, the text of a pre-print paper should vary significantly from its corresponding post-print version.

• By applying standard similarity measures, we should be able to detect and quantify such differences.
Methodology
Data Gathering

Assembling a pre-print corpus

- Source: arXiv.org
  - 1.1 million publication records
  - metadata (typical DC, including DOI) obtained via OAI-PMH interface
  - PDF versions of articles available via Amazon’s S3 service (using “requester pays” option)
  - Latest version used if multiple available
Data Gathering

Finding a matching post-print corpus

- Extract DOIs from arXiv metadata
  - 44.5% of articles have DOI

- CrossRef’s Metadata Search API
  - Match by DOI: download XML marked up or PDF full-text version of article
  - Metadata included in XML
  - Access based on UCLA’s serial subscriptions
Data Processing

Both pre-print and post-print corpora

- Convert PDF to XML where needed

- Extract sections from XML
  - Title, authors, abstract, body, references, publication date
  - Occasionally, sections are missing

- Also try extraction from arXiv’s OAI-PMH interface
  - Data provided by authors: abstracts, titles
  - Texts often have markup, complicating comparisons
Text Comparison Methods

- Length ratio
- Levenshtein ratio
- Cosine similarity
- Simhash
- Jaccard coefficient
- Sorensen similarity
Text Comparison Methods 1/3

Length ratio:
Ratio of the shorter text’s length to the longer text’s length

Example
“Four score and seven” → “Four score and seven years ago“

Four score and seven (20 characters)
Four score and seven years ago (30 characters)

Length ratio: $\frac{20}{30} = .667$
Text Comparison Methods 2/3

Levenshtein edit distance:
Number of operations (insert, delete, substitute) needed to transform one string into the other

Example
“The rose is red” → “The roose is rot”
• rose -> roose (1 insertion)
• red -> rod (1 substitution, ‘o’ for ‘e’)
• rod -> rot (1 substitution, ‘t’ for ‘d’)

Levenshtein distance: 1 + 1 + 1 = 3
Text Comparison Methods 2/3

Levenshtein edit distance:
Number of operations (insert, delete, substitute) needed to transform one string into the other

Example
“The rose is red” → “The roose is rot” (edit distance: 3)

Levenshtein ratio:

\[
\frac{\text{length of text 1} + \text{length of text 2} - \text{edit distance}}{\text{length of text 1} + \text{length of text 2}}
\]

\[
\frac{15 + 16 - 3}{15 + 16} = .903
\]
Text Comparison Methods 3/3

Cosine similarity:
Two texts have a high similarity (max value: 1) if they share many of the same words in the same proportions.

In practice, common words (“and,” “the,” “is”) are ignored, and words that are more characteristic of a given text have greater importance.

Example
“QCD sum rule approach for scalar mesons as four-quark states”
“QCD sum rule approach for the light scalar mesons as four-quark states”
Cosine similarity:
Two texts have a high similarity (max value: 1) if they share many of the same words \textit{in the same proportions}.

In practice, common words ("and," "the," "is") are ignored, and words that are more \textit{characteristic} of a given text have greater importance.

Example
qcd sum rule approach \textit{for} scalar mesons \textit{as} four quark states
qcd sum rule approach \textit{for the} light scalar mesons \textit{as} four quark states
Cosine similarity:
Two texts have a high similarity (max value: 1) if they share many of the same words in the same proportions.

In practice, common words ("and," "the," "is") are ignored, and words that are more characteristic of a given text have greater importance.

Example
qcd sum rule approach for scalar mesons as four quark states
qcd sum rule approach for the light scalar mesons as four quark states

Cosine similarity: 0.8955


How Much Does $1.7 Billion Buy?  
@farbthink, @liber8er, @peterbroadwell, @mart1nkle1n  
#CNI15F, Washington, DC, 12/15/2015

Comparison of Sections
1. INTRODUCTION

Digital libraries are called upon to organize, aggregate, and preserve non-traditional digital news collections. Rather than traditionally building side-by-side non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization. The role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections. Rather than traditionally building side-by-side non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries are called upon to organize, aggregate, and preserve non-traditional digital news collections. Rather than traditionally building side-by-side non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries have increased their investment in their efforts to acquire, describe, and provide access to non-traditional news collections. Non-traditional news collections include social media data such as Twitter, Facebook, and YouTube videos. While the role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries have increased their investment in their efforts to acquire, describe, and provide access to non-traditional news collections. Non-traditional news collections include social media data such as Twitter, Facebook, and YouTube videos. While the role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries have increased their investment in their efforts to acquire, describe, and provide access to non-traditional news collections. Non-traditional news collections include social media data such as Twitter, Facebook, and YouTube videos. While the role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries have increased their investment in their efforts to acquire, describe, and provide access to non-traditional news collections. Non-traditional news collections include social media data such as Twitter, Facebook, and YouTube videos. While the role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.

Digital libraries have increased their investment in their efforts to acquire, describe, and provide access to non-traditional news collections. Non-traditional news collections include social media data such as Twitter, Facebook, and YouTube videos. While the role of digital libraries is increasingly defined by the ability to organize, aggregate, and preserve non-traditional digital collections, digital libraries are seeking to manage these collections in a way that allows for their discovery and use. Our goal in this research effort is to develop a programmatic framework to build collections of social media news coverage that is augmented with various open source tools such as temporal analysis and visualization.
Analyzing News Events in Non-Traditional Digital Library Collections

Myra Klein
University of California, Los Angeles
kleinmyra@ucla.edu

Peter Broadwell
University of California at Los Angeles
peterbroadwell@ucla.edu

ABSTRACT

Recent advances in media recording and collection have provided a wealth of data for the library and information science (LIS) community. In this paper, we describe an online archive of news broadcasts recorded from cable television and discuss its potential for use as a non-traditional digital library. We argue that traditional analytical library tools may not be applicable to digital news media: television broadcasts and social media. Instead, we introduce the NewsScape project, an online archive of news broadcasts recorded from cable television, and explain how we will need to take into consideration when linking resources related to a particular subject. Yet television news studies have noted, for example, that the topics discussed are much less personal and perhaps more representative of thinking and the May 23rd, 2014 Isla Vista, CA killing spree.

1. INTRODUCTION

Recent advances in media recording and collection have provided a wealth of data for the library and information science (LIS) community. In this paper, we describe an online archive of news broadcasts recorded from cable television and discuss its potential for use as a non-traditional digital library. We argue that traditional analytical library tools may not be applicable to digital news media: television broadcasts and social media. Instead, we introduce the NewsScape project, an online archive of news broadcasts recorded from cable television, and explain how we will need to take into consideration when linking resources related to a particular subject. Yet television news studies have noted, for example, that the topics discussed are much less personal and perhaps more representative of thinking and the May 23rd, 2014 Isla Vista, CA killing spree.

2. RELATED WORK

The NewsScape project was inspired by the fields of natural language processing (NLP), social media analysis, and natural language understanding (NLU). NLP is a field that focuses on the interaction between human language and computers, with the goal of enabling computer systems to understand, interpret, and generate human language. Social media analysis is a field that focuses on the use of social media data to understand and analyze human behavior, with the goal of enabling organizations to make better decisions. NLU is a field that focuses on the ability of computers to understand and interpret human language, with the goal of enabling computer systems to understand, interpret, and generate human language.

3. SYSTEM ARCHITECTURE

The NewsScape system is designed to provide a comprehensive view of news events in non-traditional collections. We collect news broadcasts from cable television and social media data from Twitter, and index these resources to create a searchable database. Our system uses natural language processing techniques to extract key information from the broadcasts and social media data, and to identify relationships between entities and events.

4. RESULTS

We have applied our system to a number of news events, and have observed that the system is able to identify key information and relationships with high accuracy. For example, we have observed that the system is able to identify key information such as the location of an event, the names of individuals involved, and the topics discussed, as well as relationships such as cause-effect and similarity.

5. CONCLUSIONS

In conclusion, the NewsScape project is a promising new approach to analyzing news events in non-traditional collections. Our system is able to extract key information and identify relationships with high accuracy, and it has the potential to provide a comprehensive view of news events in non-traditional collections. We believe that this approach will be valuable for libraries and information professionals, and we encourage others to explore this field.

ACKNOWLEDGMENTS

We would like to thank the UCLA library for providing the resources and support for this project. We would also like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES


ABSTRACT
Digital libraries are called upon to organize, aggregate, and
publish born-digital news media. This role also carries
many of the traditional responsibilities of library
administrators of digitized born-digital content. How does
an information scientist manage such libraries?

Digital libraries are subject to the same challenges as
traditional libraries, such as local television markets and
international news sources. The UCLA Library has
sent a representative example of the issues entailed in the
management of such non-traditional collections. The
UCLA Library has created and is stewards a news media
archive that is becoming more prevalent due to ad-
ventions in online and on-demand media and the increas-
ing use of online and offli

Categories and Subject Descriptors
H.3.7 [Categories and Subject Descriptors]: Information
storage and retrieval systems--news, television, video,
radio, other media

1. INTRODUCTION
The role of memory organizations is increasingly defined
by organizing, aggregating, and delivering born-digital
content. As such, news libraries and archives are faced
with the task of making their "new" collections as useful
as possible to scholars. This is a challenging endeavor,
and traditional analytical library tools may not be
suitable for these non-traditional collections.

2. SOURCES
The sources in our corpus consist of the
scheduled television broadcasts of both
national and local television news programs.

3. ANALYSIS
The analysis involves applying various
programming analysis tools to the corpus.

4. RESULTS
The results are organized in the form of
categories and subject descriptors.

5. CONCLUSIONS
The conclusions are summarized in
the abstract.

References
[1] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan,
F. Zhao, X. Jiang, and B. Liu. Toward a Digital News
Library: Capturing, Storing, and Delivering Digital News
conference on Digital libraries, pages 469–477.
.com/socialfeedmanager/
Bali flights. Bali flights.
[5] F. Chierichetti, J. Kleinberg, M. Mahdian, A. Pogany,
A. Khan, J. Palau, M. Gawlick, S. Wong, and
Y. Gilfanov. Social Media Analysis: Twitter at Newsroom
abs/1204.5133

Figure 3: Sequentiality of terms in tweets and on television

4.3 Sequentiality
The sequentiality of terms that occur in both collections
we examined the sequence in which individual terms occur
across the two corpora. This allows us to identify terms
that appear on Twitter before they appear on television.

Figure 4: Comparison of Sections

References
[1] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan,
F. Zhao, X. Jiang, and B. Liu. Toward a Digital News
Library: Capturing, Storing, and Delivering Digital News
conference on Digital libraries, pages 469–477.
.com/socialfeedmanager/
Bali flights. Bali flights.
[5] F. Chierichetti, J. Kleinberg, M. Mahdian, A. Pogany,
A. Khan, J. Palau, M. Gawlick, S. Wong, and
Y. Gilfanov. Social Media Analysis: Twitter at Newsroom
abs/1204.5133

Figure 3: Sequentiality of terms in tweets and on television

4.3 Sequentiality
The sequentiality of terms that occur in both collections
we examined the sequence in which individual terms occur
across the two corpora. This allows us to identify terms
that appear on Twitter before they appear on television.

Figure 4: Comparison of Sections

References
[1] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan,
F. Zhao, X. Jiang, and B. Liu. Toward a Digital News
Library: Capturing, Storing, and Delivering Digital News
conference on Digital libraries, pages 469–477.
.com/socialfeedmanager/
Bali flights. Bali flights.
[5] F. Chierichetti, J. Kleinberg, M. Mahdian, A. Pogany,
A. Khan, J. Palau, M. Gawlick, S. Wong, and
Y. Gilfanov. Social Media Analysis: Twitter at Newsroom
abs/1204.5133

Figure 3: Sequentiality of terms in tweets and on television

4.3 Sequentiality
The sequentiality of terms that occur in both collections
we examined the sequence in which individual terms occur
across the two corpora. This allows us to identify terms
that appear on Twitter before they appear on television.

Figure 4: Comparison of Sections

References
[1] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan,
F. Zhao, X. Jiang, and B. Liu. Toward a Digital News
Library: Capturing, Storing, and Delivering Digital News
conference on Digital libraries, pages 469–477.
.com/socialfeedmanager/
Bali flights. Bali flights.
[5] F. Chierichetti, J. Kleinberg, M. Mahdian, A. Pogany,
A. Khan, J. Palau, M. Gawlick, S. Wong, and
Y. Gilfanov. Social Media Analysis: Twitter at Newsroom
abs/1204.5133

Figure 3: Sequentiality of terms in tweets and on television

4.3 Sequentiality
The sequentiality of terms that occur in both collections
we examined the sequence in which individual terms occur
across the two corpora. This allows us to identify terms
that appear on Twitter before they appear on television.
Preliminary Findings
Data Gathering Results

Both pre-print and post-print corpora

- 11,017 full text articles matched by DOI
- Most papers within date range 2003 – 2015
- 96% of post-prints published by Elsevier
- “Physics Letters B” top journal
Title Comparison

Browse findings at http://sologlo.library.ucla.edu/prepost
Abstract Comparison

Browse findings at http://sologlo.library.ucla.edu/prepost

How Much Does $1.7 Billion Buy?
@farbthink, @liber8er, @peterbroadwell, @mart1nkle1n
#CNI15F, Washington, DC, 12/15/2015
Body Comparison

Browse findings at http://sologlo.library.ucla.edu/prepost
One of the frontiers of QCD which are intensely investigated in high energy experiments is the high energy (small x) regime, where we expect to observe the non-linear behavior of the theory. In this regime, the growth of the parton distribution should saturate, forming a Color Glass Condensate (CGC). In fact, signals of parton saturation have already been observed both in ep deep inelastic scattering at HERA and in deuteron-gold collisions at RHIC. Currently, a global description of the existing experimental data is possible considering different phenomenological saturation models for the two processes within the CGC formalism. In this Letter we analyze the universality of these dipole cross section parameterizations and verify that they are not able to describe the HERA and RHIC data simultaneously. We analyze possible improvements in the parameterizations and propose a new parameterization for the forward dipole amplitude which allows us to describe quite well the small-x ep HERA data on F2 structure function as well as the dAu RHIC data on charged hadron spectra. It is an important signature of the universality of the saturation physics.

Abstract

One of the frontiers of QCD which are intensely investigated in high energy experiments is the high energy (small x) regime, where we expect to observe the non-linear behavior of the theory. In this regime, the growth of the parton distribution should saturate, forming a color glass condensate (CGC). In fact, signals of parton saturation have already been observed both in ep deep inelastic scattering at HERA and in deuteron-gold collisions at RHIC. Currently, a global description of the existing experimental data is possible considering different phenomenological saturation models for the two processes within the CGC formalism. In this Letter we analyze the universality of these dipole cross section parameterizations and verify that they are not able to describe the HERA and RHIC data simultaneously. We analyze possible improvements in the parameterizations and propose a new parameterization for the forward dipole amplitude which allows us to describe quite well the small- x ep HERA data on F2 structure function as well as the dAu RHIC data on charged hadron spectra. It is an important signature of the universality of the saturation physics.
Author Similarity (very preliminary)

• Not trivial to determine due to
  • Different name formatting
  • Varying metadata quality

“Save” statement: 7,731/9,935 articles (78%) have identical author lists (number and ranks of authors)
Next Steps
More to Come!

- Refine extraction/comparison of authors and references

- Overlay with ISI Impact factor and usage statistics

- Expand to other disciplines/publishers
  - Social Sciences
  - Humanities
  - Economy
  - Linguistics

- Operate at scale
  - Seeking collaboration
  - Enable other institutions to conduct similar experiments
Discussion
Questions

• Collecting imperative
  • New forms of scholarly communication
  • Developing data collections

• UC faculty is contributing to 1/12 of Elsevier publications, yet we license back 100%

• Open Access
  • Not even APCs added to the calculations
  • Revenue: $182 million in 2012, growing 30% per year

• What about preservation?
How Much Does $1.7 Billion Buy?

A Comparison of Scientific Journal Articles to Their Pre-print Versions

Sharon E. Farb @farbthink
Peter Broadwell @peterbroadwell
Martin Klein @mart1nkle1n
Todd R. Grappone @liber8er
If Harvard Can’t Afford It, Who Can?

Harvard University says it can't afford journal publishers' prices

University wants scientists to make their research open access and resign from publications that keep articles behind paywalls

A graduation ceremony at Harvard University. Photograph: Brooks Kraft/Corbis
Increase in Global Scientific Output

Global scientific output doubles every nine years

07 May 2014 11:46 GMT | Posted by Richard Van Noorden | Category: Policy, Publishing

It’s a common complaint among academicians: today’s researchers are publishing too much, too fast. But just how fast is the mass of scientific output actually growing?

Many would throw up their hands and declare the question impossible. It’s clearly wrong to cite the growth of academic databases, such as Thomson Reuters Web of Science, which has increased its coverage by around 3% per year (barring occasions when the database incorporates a flood of new journals). That dramatically undercounts the true expansion: no database captures everything.

Bibliometric analysts Lutz Bornmann, at the Max Planck Society in Munich, Germany and Ruediger Mutz, at the Swiss Federal Institute of Technology in Zurich, think they have a better answer. It is impossible to know for sure, but the real rate is closer to 6-8% each year, they argue. That equates to a doubling of global scientific output roughly every nine years.
Maybe It’s Time for an Open Access Model

If Harvard Can’t Afford Academic Journal Subscriptions, Maybe It’s Time for an Open Access Model

By Keith Wagstaff @kwagstaff | April 26, 2012 | Add a Comment

Last week, Harvard’s Faculty Advisory Council revealed that the school now spends $3.75 million annually on academic journal subscriptions. Why so much? According to a memo the council sent out, some journals cost the school up to $40,000 every year, with the two top publishers increasing the price of content 145% over the last six years.

This is troubling for a number of reasons. First, in an age where the public can browse nearly 4 million articles for free on Wikipedia, a curious person looking to read up on the latest scientific research can expect to spend nearly $30 to $40 for a single paper from publishers such as Elsevier and Springer.
Begin with $1 Billion Cost Per Year with a 6% Increase Annually

[VALUE]

$2,012,196,472

$1,898,298,558

$1,790,847,697

$1,689,478,959

$1,593,848,075

$1,503,630,259

$1,418,519,112

$1,338,225,578

$1,262,476,960

$1,191,016,000

$1,123,600,000

$1,060,000,000

$120,731,788

$113,897,914

$60,000,000 $63,600,000 $67,416,000

$71,460,960 $75,748,618 $80,293,535 $85,111,147 $90,217,816 $95,630,884 $101,368,738 $107,450,862


Amount of Increase

Total

How Much Does $1.7 Billion Buy?
@farbthink, @liber8er, @peterbroadwell, @mart1nkle1n
#CNI15F, Washington, DC, 12/15/2015
How Much Does $1.7 Billion Buy?

@farbthink, @liber8er, @peterbroadwell, @mart1nkle1n
#CNI15F, Washington, DC, 12/15/2015

Profit Margin

Data from 2007 or 2008.
*Adjusted Operating Margin
**RMA Annual Statement Studies, 2007

Data Source: MIT Libraries
The Rise of OA “Overlay” Journals?

Leading mathematician launches arXiv 'overlay' journal

Journal that reviews papers from preprint server aims to return publishing to the hands of academics.

Philip Ball

15 September 2015

New journals spring up with overwhelming, almost tiresome, frequency these days. But Discrete Analysis is different. This journal is online only — but it will contain no papers. Rather, it will provide links to mathematics papers hosted on the preprint server arXiv. Researchers will submit their papers directly from arXiv to the journal, which will evaluate them by conventional peer review.