A Grounded Theory of Information Quality for Web Archives

Brenda Reyes Ayala[1]

[1] University of North Texas

Dissertation Defense, May 18, 2018

Committee Members:
Dr. Jiangping Chen (Co-chair)
Dr. Oksana Zavalina (Co-chair)
Dr. Shawne Miksa
Dr. Kathryn Masten-Cain

Overview I

1. Introduction

- Types and Severity of Quality Problems

2. Problem Statement

3. Purpose

4. Research Questions

5. Contributions to the Research Area

6. Literature Review

- IQ in Information Science
- IQ in Computer Science
- Research on Quality in Web Archives

7. Methodology, Data Collection, Analysis

Overview II

- Methodology
- Methodological Issues
- Data Collection and Analysis

8. Scope and Limitations

9. Results: A Grounded Theory of IQ for Web Archives

- Correspondence
- Relevance
- Archivability

10. Discussion and Conclusions

- Unexpected or Surprising Findings
- Future Work

Overview III

11. References

Terms and Definitions

- Web Archiving: the action of storing websites to preserve them as a historical, informational, legal, or evidential record
- Web Archive: a system that contains such records

Example of a web archive:
Climate change and environmental policy web archive

The Digital Dark Age

In a conference for the International Federation of Library Associations and Institutions, Terry Kuny referred to the future as a “Digital Dark Age,” an “era where much of what we know today, much of what is coded and written electronically, will be lost forever” (Kuny, 1997).

Timeline of Web Archiving

1996 Internet Archive is founded with the mission of creating a universally accessible digital library. National Library of Australia inaugurates the first-ever web archiving program by a national library

2000 Library of Congress began its Minerva Project (now the Library of Congress Web Archives)

2003 International Internet Preservation Consortium (IIPC) is founded with the mission of "improving the tools, standards and best practices of web archiving while promoting international collaboration, broad access and use of web archives for research and cultural heritage"

2004 British Library launches the UK Web Archive

2006 National Library of France launches the French Web Archive

Web Archiving in the American Context

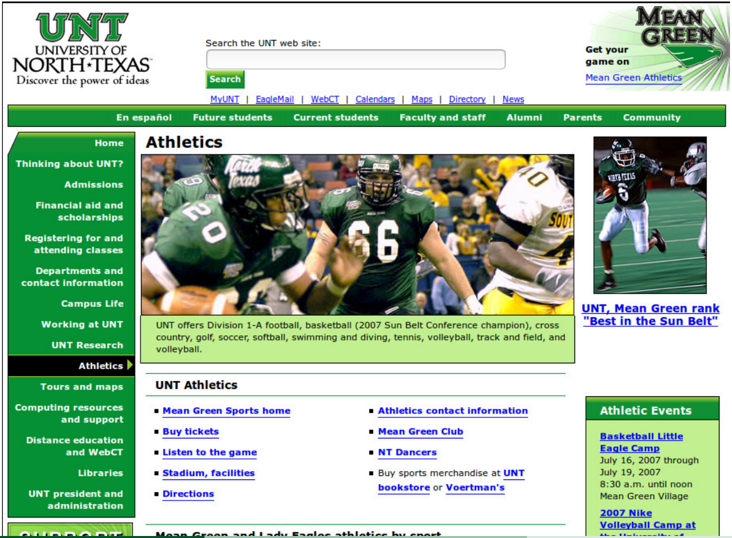
2016 Survey on web archiving in the United States, conducted by the NDSA. Over 100 American institutions that had web archiving programs in place:

- 63% colleges and universities
- 15% federal, state, and local governments
- 12% archives

(Bailey, Grotke, McCain, Moffatt, & Taylor, 2016).

Good-Quality Archived Website

A good-quality archived version of the UNT Athletics site from 2007



Search the UNT web site:

Get your game on

Mean Green Athletics

MyUNT | EagleMail | WebCT | Calendars | Maps | Directory | News

En español | Future students | Current students | Faculty and staff | Alumni | Parents | Community

- Home
- Thinking about UNT?
- Admissions
- Financial aid and scholarships
- Registering for and attending classes
- Departments and contact information
- Campus Life
- Working at UNT
- UNT Research
- Athletics
- Tours and maps
- Computing resources and support
- Distance education and WebCT
- Libraries
- UNT president and administration

Athletics

UNT offers Division 1-A football, basketball (2007 Sun Belt Conference champion), cross country, golf, soccer, softball, swimming and diving, tennis, volleyball, track and field, and volleyball.

UNT, Mean Green rank "Best in the Sun Belt"

UNT Athletics

- Mean Green Sports home
- Buy tickets
- Listen to the game
- Stadium, facilities
- Directions
- Athletics contact information
- Mean Green Club
- NT Dancers
- Buy sports merchandise at UNT bookstore or Voertman's

Athletic Events

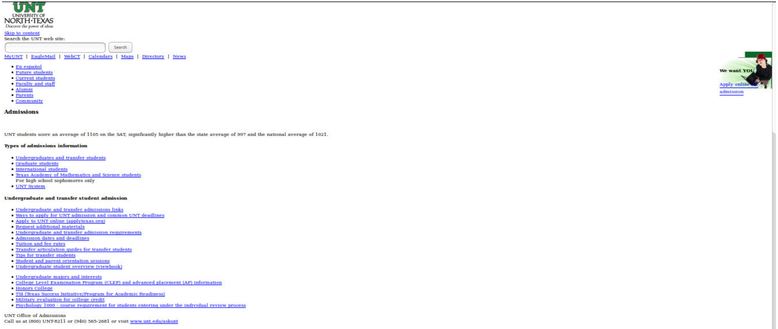
Basketball Little Eagle Camp
July 16, 2007 through July 19, 2007
8:30 a.m. until noon
Mean Green Village

2007 Nike Volleyball Camp at the University of

Mean Green and Lady Eagles athletics by sport

Medium-Quality Archived Website

An archived version of the UNT Admissions site from 2007, missing the styling of the original



UNT UNIVERSITY OF NORTH·TEXAS Discover the power of ideas

Skip to content
Search the UNT web site:

MyUNT | EagleMail | WebCT | Calendars | Maps | Directory | News

- En español
- Future students
- Current students
- Faculty and staff
- Alumni
- Parents
- Community

Admissions

UNT students score an average of 1105 on the SAT, significantly higher than the state average of 997 and the national average of 1021.

Types of admissions information

- Undergraduates and transfer students
- Graduate students
- International students
- Texas Academy of Mathematics and Science students
  For high school sophomores only
- UNT System

Undergraduate and transfer student admission

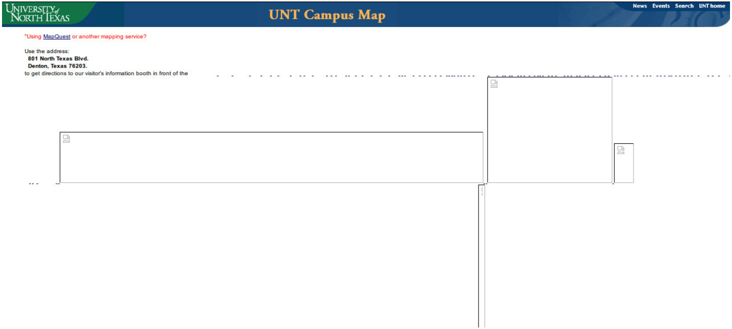
- Undergraduate and transfer admissions links
- Ways to apply for UNT admission and common UNT deadlines
- Apply to UNT online (applytexas.org)
- Request additional materials
- Undergraduate and transfer admission requirements
- Admission dates and deadlines
- Tuition and fee rates
- Transfer articulation guides for transfer students
- Tips for transfer students
- Student and parent orientation sessions
- Undergraduate student overview (viewbook)

- Undergraduate majors and interests
- College Level Examination Program (CLEP) and advanced placement (AP) information
- Honors College
- TSI (Texas Success Initiative/Program for Academic Readiness)
- Military evaluation for college credit
- Psychology 1000 - course requirement for students entering under the individual review process

UNT Office of Admissions
Call us at (800) UNT-8211 or (940) 565-2681 or visit www.unt.edu/askunt

Low-Quality Archived Website

An archived version of the UNT Campus Map from 2004, almost unusable



"Using MapQuest or another mapping service?

Use the address:
801 North Texas Blvd.
Denton, Texas 76203.
to get directions to our visitor's information booth in front of the

Problematic Definition of IQ in Web Archives

In the field of web archiving, there has been only one definition of Information Quality (IQ) in a web archive, put forward by Masanés. Quality in a web archive is made up of the following elements:

1. the completeness of material (linked files) archived within a target perimeter
2. the ability to render the original form of the site, particularly regarding navigation and interaction with the user

(Masanés, p.39)

Definition is too centered on the technological tools needed to archive websites, and is not grounded on research with humans.
System-centered and not human-centered

Lack of Models, Theories or Frameworks for Web Archiving

The technical developments in the field have far outpaced the development of proper theoretical tools or models. Over two decades into its history, web archiving still lacks a theoretical underpinning. Essentially, we have technological tools to build web archives, but no conceptual tools to understand them

Purpose

To build a theory of IQ for web archives that is grounded in human-centered data

Research Questions

RQ 1

What is the definition of information quality (IQ) for web archives?

RQ2

How can IQ in a web archive be measured?

Outline

A study that is both descriptive and experimental.

Phase 1: (RQ 1)

Use a Grounded Theory (GT) approach to create a theory of IQ for web archives. Iteratively refine it

Phase 2 (RQ 2)

Use the theory to determine how IQ dimensions can be operationalized

Significance

1. Begin the work of establishing a much-needed theoretical groundwork for the field, which will help its development and growth
2. Allow practitioners in the field of web archiving to apply it to measure the quality of their own web archives and improve the Quality Assurance processes for their organizations

Taylor: Value Added Model

1. Accuracy: a guarantee of a true copy, but is independent of the truth value of the information.
2. Comprehensiveness: the completeness of coverage of a particular subject or discipline.
3. Currency: recency of the data acquired by the system and the capability of the system to reflect current modes of thinking in its access vocabularies.
4. Validity:degree to which the information or data presented to users can be judged as sound.
5. Reliability: the trust a user has in the consistency of quality performance of the systems and its outputs over time. Taylor states that reliability is the summation of many aspects of quality.

(Taylor, 1986)

Rieh: Five Aspects of Information Quality on the Web

1. Goodness: Good job, bad, better, excellent, fine, nice, great, best, perfect,
2. Accuracy: Accurate, correct, right, precise
3. Currency: Current, recent, up-to-date, out-of-date, old, timely
4. Usefulness: Useful, useless, hard to use, informative, helpful, doesn't help,
5. Importance: Important

(Rieh, 2002)

Stvilia's Framework for Information Quality

Stvilia (2006) developed a general IQ measurement and assessment framework based on study of two large-scale collections: Simple Dublin Core (DC) metadata records and Wikipedia articles

Three high-level categories of IQ:

1. Intrinsic: dimensions of IQ assessed by measuring internal attributes/characteristics of information entities themselves in relation to some reference standard in a given culture. Intrinsic IQ attributes persist and depend little on context
2. Relational/Contextual: measures relationships between information and some aspects of its usage context. These characteristics are not persistent with the entity itself
3. Reputational: measures the position of an information entity in a cultural or activity structure, often determined by its origin and its record of mediation

Stvilia's Dimensions of Quality I

Definitions given are for relational/contextual IQ dimensions

1. Accuracy/Validity: degree to which information object correctly represents another information object, process or phenomenon in the context of a particular activity/culture
2. Accessibility: speed, ease of locating and obtaining an information object relative to particular activity
3. Complexity: degree of cognitive complexity of information object relative to particular activity
4. Informativeness/Redundancy: extent to which information is new or informative in the context of particular activity/community
5. Naturalness: degree to which information objects model and content are semantically close to the objects, states or processes they represent in the context of particular activity

Stvilia's Dimensions of Quality II

6. Precision/completeness: extent to which information object matches the precision and completeness needed in the context of given activity
7. Relevance (aboutness): extent to which information is applicable and helpful/applicable in given activity
8. Security: extent of protection of information from harm in the context of a particular activity
9. Semantic consistency: extent of consistency of using the same values (vocabulary control) and elements required or suggested by some external standards
10. Structural consistency: extent to which similar attributes or elements of information object are consistently represented with the same structure, format and precision required

Stvilia's Dimensions of Quality III

11. Verifiability: extent to which the correctness of information is verifiable and/or provable in the context of a particular activity
12. Volatility: amount of time the information remains valid in the context of a particular activity

Stvilia's Dimensions of Quality, Operationalized

Dimension	Metric
Accuracy/Validity	Num. Broken External Links; Euclidean Similarity Distance
Accessibility	Throughput w/r of # Requests; Throughput w/r of Amount of data
Complexity	Flesch Reading Ease Score; Flesch- Kincaid Grade Level; Fog Index; Average Sentence Length; Average Word Length
Informativeness	Kullback-Leibler Divergence; IDF/AverageIDF
Naturalness	Cosine Angular Similarity Metric
Precision/ completeness	Completeness Ratio; FRBR Completeness

Stvilia's Dimensions of Quality, Operationalized II

Dimension	**Metric**
Relevance	NumberOfClicks; CitationCount; VectorSpaceModel; AuthorityAndHub; PageRank
Security	Break-ins Ratio
Semantic consistency	Semantic Consistency Index
Structural consistency	Structural Consistency Index
Verifiability	Num. of URLs; Num. of References (citing)
Volatility	LinkRot

Zhu and Gauch: Quality Metrics Appropriate for Automatic Analysis

- Currency: how recently a web page has been updated. The time stamp of the last modification of the document
- Availability: the number of broken links contained by the web page. The number of broken links on a page divided by the total numbers of links it contains
- Information-to-Noise Ratio: proportion of useful information contained in a web page of a given size. The total length of the tokens divided by the size of the document.
- Authority: the reputation of the organization that produced the web page. Based on the Yahoo Internet Life (YIL) reviews

Zhu and Gauch: Quality Metrics Appropriate for Automatic Analysis II

- Popularity: how many other Web pages have cited this particular Web page. The number of links pointing to a web page
- Cohesiveness: the degree to which the content of the page is focused on one topic. How closely related the major topics in the page are

(Zhu & Gauch, 2000, p.289)

Zhu and Gauch: How to Calculate IQ

The “goodness” of a site:

$$G_i = \overline{W}_i * (a''_s * \overline{T}_i + b''_s * \overline{A}_i + c''_s * \bar{I}_i + d''_s * \overline{R}_i + e''_s * \overline{P}_i + f''_s * \overline{C}_i)$$

where $\overline{W}_i, \overline{T}_i, \overline{A}_i, \bar{I}_i, \overline{R}_i, \overline{P}_i$ are the means of information quantity, currency, availability, information-to-noise ratio, authority, and popularity of site i across topics relevant to the query, $\overline{C}_i$, is the cohesiveness of site *i*, $a''_s, b''_s, c''_s, d''_s, e''_s, f''_s$ are the weights representing the importance of each quality metric. (Zhu & Gauch, 2000, p.291)

The Notion of Archivability

- Banos and Manolopoulos (2015) introduced website archivability: “sum of the attributes that make a website amenable to being archived” (p.1)
- Archivability depends on five facets: accessibility F_A, standards compliance F_S, cohesion F_C, and metadata usage F_M
- Website has a score for each facet, represented as a tuple. The value of x_k is either 0 or 1, which represents negative or positive answer to a specific criterion
- Components of a single facet are assigned a weight (ω_k) depending on their significance. For high-significance components, $\omega_k = 4$, $\omega_k = 2$ for medium-significance components and $\omega_k = 1$ for low-significance components

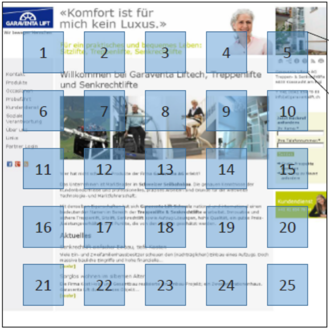
$$WA = \sum_{\lambda \in \{A,S,C,M\}} w_\lambda F_\lambda$$

Quality and Similarity at the Swiss National Library

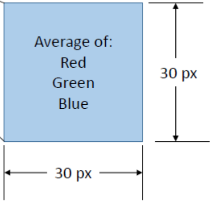
- In 2014, the Swiss National Library (NL) began implementing a system called the Visual Quality Indicator (VQI) to help their web archivists better conduct QA
- VQI system uses Euclidean distance metric to compare the visual appearance of an archived website in an ongoing crawl to A) the live website and B) the most recent version of the archived website stored in the library
- The greater the distance, the greater the difference between the two images, and thus, the greater the difference between the two websites
- If the distance is high enough, web archivists will perform QA on the archived site

Quality and Similarity at the Swiss National Library II

VQI calculation – Calculation of region vectors


«Komfort ist für mich kein Luxus.»
Willkommen bei Garaventa Liftech, Treppenlifte und Senkrechtlifte
Aktuelles
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Each screenshot is divided into 25 individual regions, aligning the parts by 5 x 5. For each region, the average value of the contained RGB values is calculated.



Grounded Theory Methodology (GT)

- Introduced by Barney Glaser and Anselm Strauss in their 1967 book *The Discovery of Grounded Theory*
- The discovery of theory from data, systematically obtained and analyzed
- Conceived as a reaction to then-prevalent trends; increased emphasis on verification had caused a dearth of theories in the field of sociology
- In the decades that followed, GT became a major methodology used by researchers in a wide variety of fields. It has been further clarified by Glaser and Strauss, and added to by researchers such as Kathy Charmaz, Juliet Corbin, and Adele Clarke

Differences between GT and Logico-formal Theory

Characteristic	Traditional Approach	Grounded Theory
Literature Review	Before data collection	Throughout data collection, analysis
Method	Compare only “comparable" groups	Compare any groups
Sampling	Statistical sampling	Theoretical sampling
Data	Field notes, interviews, observations	Wide variety of materials
Data Collection	After theory is formulated	At any time
Purpose	To verify theory	To generate theory
Goal	To establish fact	To establish structural boundaries of fact
View theory as	A perfected product	An ever-developing entity

Lazarsfeld's Qualitative Mathematics

1. Created by sociologist Paul L. Lazarsfeld, who pioneered the idea of using mathematical reasoning as an aid to theory building in the social sciences
2. Based on the premise that using mathematics does not lead to new findings, but it can clarify relationships
3. Key concepts
   - Variate/indicator: any classificatory or ordering device by means of which distinctions can be made among people or collectives
   - Variates are combined to form an *index*. A theory, whether mathematical or not, is meant to express relationships between indices.
   - Interchangeability of indices

Methodological Issues, Auditing the Dissertation

Purposeful peer review:

- Committee members were periodically invited to audit the entire research project, including the codebook, preliminary findings, and core categories. Employees of the Internet Archive were also invited to see the findings
- Summer of 2017, the researcher presented the theory and her preliminary findings at the doctoral consortium of the Joint Conference of Digital Libraries, where she received feedback from her peers and other IS researchers

Methodological Issues, Lazarsfeld's Panel Analysis

Original intent to use Lazarsfeld's panel analysis method to operationalize the dimensions of quality was not entirely successful

- Panel analysis approach makes heavy use of time as an important variable to take into account
- While time is indeed an important variable that should be considered for web archives, it did not arise as a category (indeed it rarely surfaced) in the data
- According to the rules of CGT, a category cannot be created if it is not present in the data, and thus it was not included in the final theory
- Despite these issues, Lazarsfeld's general principles of qualitative mathematics proved invaluable while operationalizing the dimensions of quality

Phase 1: Building a Substantive Theory of Quality in a Web Archive

The Internet Archive's Archive-It (AIT) is a subscription-based web archiving service that helps organizations build and manage their own web archives.

1. Negotiated a researcher agreement with the Internet Archive to obtain a large cache of AIT support tickets
2. Cleaned the data using several Python programs and Linux command-line scripts that I created. Imported the clean data into Nvivo software package
3. Used open coding and theoretical memos to identify the main concepts and categories present in the data
4. Created a preliminary theory of IQ for web archives
5. Used the constant comparison method to refine and improve the theory
6. Performed literature review

Sample AIT Support Ticket I

AIT client

In the collection, there seems to be a problem with the way the _____.com is being captured–when I try to access it through Wayback, all I get is a blank page (see attached). Does the way the _____.com site is built that makes it uncapturable? I did a test run before adding it to our collection, and don't remember this being a problem. Thanks!

Sample AIT Support Ticket II

AIT employee

Hi _____, It looks like most of the content from the _____.com site has been captured, but display is a bit tricky due to the extensive use of javascript and flash on this site. For now, you can view the archived content for this site by disabling javascript in your browser while you're browsing through the site in wayback.

It may not have the exact look and feel of this site on the live web, but you should be able to see most of the archived content. I will also send this along to our engineers to look into a more long-term solution to allow users to more easily view this content, and we'll update you when we have more information.

Please let me know if you have any further questions!

Sample AIT Support Ticket III

AIT client

Thanks _____! I can see most of the content when I disable Javascript as you suggest–but would love a more long-term solution! I'll wait to hear if/when more information is available.

Phase 2: Identifying the Operationalizable Dimensions of Web Archive Quality

1. Explored the relationships between aspects of IQ
2. Constructed operational definitions of each IQ dimension
3. Performed additional literature review
4. Further refined and modified the operational definitions

Scope and Limitations

- Not all dimensions of IQ can be easily measured quantitatively. Some IQ dimensions such as “usefulness” are impossible to measure because they depend entirely on the user’s opinion
- Some measures of IQ, such as correspondence, might not be available. For example, if the original site has been lost, there is no way to compare it to the archived version
- Focusing on only those dimensions of IQ that can be operationalized would limit the scope of the theory

Scope and Limitations of the Theory

- This is a *substantive* theory, that is, it is specific to the context of web archiving and not meant to describe the construct of IQ in a more general form
- The theory is delimited because it is specific to small or medium-size web archives that are focused on covering a single topic or an event

Scope and Limitations, Assumptions

- Countable and finite: the original and archived websites form a finite, countable set
- Closed World Assumption (CWA): all the elements of a website are represented in the components, c_n, of the set O or c'_n of the set A
- Web page vs. website vs. web archive: the theory focuses on IQ at the webpage level; however, some dimensions such as topic and size relevance are more appropriately measured at the website and web archive level

A Look at the Data Analyzed

Number of Tickets and Interactions About Information Quality Analyzed Per Year

Year	No. IQ tickets analyzed	No. interactions analyzed
2012	74	478
2013	65	492
2014	67	540
2015	58	528
2016	41	506
Total	305	2544

Core Facets of IQ for Web Archives

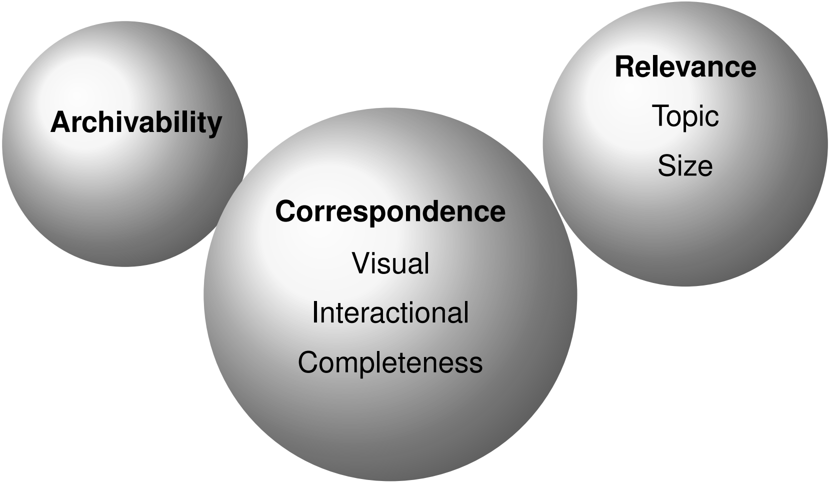
1. **Correspondence**: similarity between the original and archived websites. Good correspondence requires equivalence, or at least a close resemblance, between the two
   - Visual
   - Interactional
   - Completeness
2. **Relevance**: pertinence of the contents of an archived website to the original. Archived websites must not contain off-topic content (topic relevance) or content in quantity or volume that is unexpected or excessive (size relevance)
   - Topic
   - Size
3. **Archivability**: intrinsic properties of a website that make it more difficult to archive. A latent IQ dimension

Dimensions of IQ in a Web Archive and their Frequencies

Counts in the sub-categories do not add up to the totals in their main categories because many interactions were coded as belonging to more than one core category and NVivo counts them as such.

Dimension	No. of Mentions	No. of Tickets
Correspondence	852	226
Visual correspondence	160	91
Interactional correspondence	72	49
Completeness	478	157
Relevance	451	127
Topic relevance	93	54
Size relevance	351	107
Archivability	101	78

The Theory of IQ for Web archives, Visualized



Archivability

Correspondence

Visual

Interactional

Completeness

Relevance

Topic

Size

Caveats of Operationalization

- Lazarsfeld notion of the interchangeability of indices: substituting one index for another, or adding additional indices to the formula is unlike to change the direction of the general relationship
- As a result, the researcher was able to pick a similarity measure (say, cosine similarity) to explain an IQ dimension, but was also able to say that other measures such as Jaccard similarity or Euclidean distance would also prove useful
- Overall, Lazarsfeld's methods, while not followed to the letter by the researcher, provided flexible principles that enabled creative and adaptable ways of operationalizing quality

Correspondence

- Visual: similarity in appearance between the original website and the archived website
- Interactional: degree to which a user's interaction with the archived website is similar to that of the original
- Completeness: degree to which the archived website contains all of the components of the original

Examples of Visual Correspondence

Example 1

“On the new http://www.stateu.edu/academics page we are not capturing the background images. I cannot figure out why since we are capturing other images from the same directory”

Example 2

“One thing related though, the page is not capturing its look and feel well... Any suggestions? It’s missing the background and objects are not in the right locations”

Example 3

“We’re having some trouble with our Facebook site captures not displaying properly (or at all, really)”

Operationalizing Visual Correspondence

1. Create screenshots of the original website and the archived website
2. Divide each screenshot into 25 different regions
3. Calculate the average RGB values for each region
4. Calculate the Euclidean distance between the RGB values of each screenshot. The greater the distance, the greater the difference between the two images, and thus, the greater the difference between the two websites

Operationalizing Visual Correspondence II

Visual correspondence is inversely proportional to the Euclidean distance between the original and archived webpages.

$$VC(O,A) = \frac{1}{ed(O,A)} = \frac{1}{\sqrt{\sum_{i=1}^{N} (c_i - c'_i)^2}}$$

A high value of Euclidean distance would indicate greater differences between the original and archived websites, and thus a lower degree of visual correspondence.

Examples of Interactional Correspondence

Example 1

“the interactive floorplan isn’t working as it should do - the text should appear over the map when you click on it, rather than in a list underneath”

Example 2

“When i click on it, it briefly flashes to the homepage and then it displays a URL with the nationalscience URL in it twice”

Example 3

“I would like to know if there is any way I can capture the search feature of the website, which is with the search box on the top right of the site attached http://mishima.jp/”

Operationalizing Interactional Correspondence, Network Requests

Sample of Network Requests that Generated Errors for the Archived Version of the UNT Campus Map

File Name	Cause	Live	Archived
StaticMapService. GetMapImage	img	200	404
ViewportInfoService. GetViewportInfo	script	200	404
AuthenticationService. Authenticate	script	200	404
ga.js	script	200	404

Operationalizing Interactional Correspondence, Definitions

Let us define the following sets:

Definition

1. N_A is the set of network requests of the archived website
2. N_O is the set of network requests of the original website
3. $N_O \cap N_A$ (or N_{OA})is the set of requests common to both the original and the archived websites. Only this set is of interest
4. $N_E = \{x : x \in (N_O \cap N_A)$ and x has caused an error in $N_A\}$
5. $N'{}_E = \{x : x \in (N_O \cap N_A)$ and x has not caused an error in $N_A\}$
6. $N_E \cup N'_E = N_O \cap N_A$

Operationalizing Interactional Correspondence, Formula

$$IC = \frac{|N'_E|}{|N_O \cap N_A|}$$

Example

$|N_O| = 100$, $|N_A| = 120$
$|N_O \cap N_A| = 90$
$|N_E| = 20$ and $|N'_E| = 70$
Using the formula IC is equal to 70/90 or $0.\overline{7}$. Therefore the archived website has 77.78 % of the interactional correspondence of the original

Operationalizing Interactional Correspondence, Weighted Definitions

Use a weighted version that takes into account the importance of a specific component

Definition

The set N_E can be defined as $N_E = \{x : x \in (N_{E_H} \cup N_{E_M} \cup N_{E_L})\}$, where

1. N_{E_H} is the set of network requests in both O and A that caused errors *of high importance* in the archived website; $w_i = 4$
2. N_{E_M} is the set of network requests in both O and A that caused errors *of medium importance* in the archived website; $w_i = 2$
3. N_{E_L} is the set of network requests in both O and A that caused errors *of low importance* in the archived website; $w_i = 1$
4. $N_{E_H} \cap N_{E_M} \cap N_{E_L} \equiv 0$, that is, the three sets are disjoint

Operationalizing Interactional Correspondence, Formula for Weighted Version

$$IC = \frac{N'_E}{N_O \cap N_A} = \frac{w_i * |N'_{E_H}| + w_i * |N'_{E_M}| + w_i * |N'_{E_L}|}{w_i * |N_{OA_H}| + w_i * |N_{OA_M}| + w_i * |N_{OA_L}}$$

Example

$|N_{OA}| = 150$, $|N_E| = 50$ and $|N'_E| = 100$
$|N_{OA_H}| = 55$, $|N_{OA_M}| = 55$, and $|N_{OA_L}| = 40$
$|N_{E_H}| = 35$, $|N_{E_M}| = 12$, and $|N_{E_L}| = 3$
$|N'_{E_H}| = 20$, $|N'_{E_M}| = 43$, and $|N'_{E_L}| = 37$
weighted IC = $\frac{35(4) + 12(2) + 3(1)}{55(4) + 55(2) + 40(1)} = \frac{167}{370} = 0.45$

Examples of Completeness

Example 1

“on all most every blog that we have captured from blogspot the Wayback Machine does not include the subsequent pages beyond the first”

Example 2

“It looks like in this case, the ‘Older Posts’ page was not captured because it was blocked by robots.txt”

Example 3

“The News pages (which are located under each individual sport) are being captured, but the actual articles that are listed and linked out are not”

Operationalizing Completeness, Cosine Similarity

We can operationalize completeness as the cosine similarity between *o*, the original website and *a*, the archived website.
If we express *o* and *a* as bit vectors *O* and *A* that contain all the components, *c*, of a website, such as text, images, video, etc, then the cosine similarity becomes:

$$cosine(O, A) = \frac{O \cdot A}{\parallel O \parallel \parallel A \parallel} == \frac{\sum_{i=1}^{n} c_i * c'_i}{\sqrt{\sum_{i=1}^{n} {c_i}^2} * \sqrt{\sum_{i=1}^{n} {c'_i}^2}}$$

$$O =< c_1, c_2, c_3, c_4, c_{5,}, ...c_n >$$
$$A =< c'_1, c'_2, c'_3, c'_4, c'_{5,}, ...c'_n >$$

Operationalizing Completeness, Assumptions

- In cosine similarity, the values of a vector can be binary, that is, 0 or 1. Let us assume that the value of each component, c, is also binary. So $c_n = 0$ if the component is absent, and $c_n = 1$ if the component is present.
- Let us assume that the original website, O, always has all of its components, so $O =< 1, 1, 1, 1, 1, ...1 >$
- The archived website, $A =< c'_1, c'_2, c'_3, c'_4, c'_{5,}, ...c'_n >$, since we do not yet know the values of A.

Operationalizing Completeness, Substitution

Then we can proceed to calculate the magnitudes of the original site and the archived site:

$$\parallel O \parallel = \sqrt{\sum_{i=1}^{n} c_i{}^2} = \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n}$$

$$\parallel A \parallel = \sqrt{\sum_{i=1}^{n} c_i'^2} = \sqrt{c_1'{}^2 + c_2'{}^2 ... + c_n'{}^2}$$

As well as their dot product:

$$O \cdot A =< 1, 1, ..1 > \cdot < c_1', c_2', ..c_n' = \sum_{i=1}^{n} (1) c_i' = \sum_{i=1}^{n} c_i'$$

Operationalizing Completeness, Generalization

Substituting these values into the equation, we get the following, generalized version of completeness:

$$cosine(O, A) = \frac{\sum_{i=1}^{n} c_i'}{\sqrt{n * \sum_{i=1}^{n} {c_i'}^2}}$$

Relevance

- Topic: degree to which an archived website (or a web archive) includes only content that is closely related to that of the original website or the topic of the larger web archive
- Size: similarity in size of the archived website to the original website

Examples of Problems with Topic Relevance

Example 1

“The problem is, that a lot of *unrelated* content is being displayed: sites we are not supposed to have in our collection, social network pages like xing and facebook,porn and dating sites, some of them even with illegal content, and so on”

Example 2

“the seed http://www.oakschools.org has tons of *garbage* URLs”

Example 3

“I noticed that we captured a message board that has a lot of *unwanted garbage* posted on it”

Examples of Problems with Size Relevance

Example 1

“The crawl took 12 hours and returned 103,173 documents and 3.1GB of data. This can not be correct. Crawling the whole law.stateu.edu domain with my contraints yields 20,300 +- docs”

Example 2

“There are only 170 photos on this site but I ended up with 15K new URLs”

Example 3

“There were more than 300,000 URLs queued when my time limit ran out! Looking through the queued URLs, it looks like this site is using some jQuery tools (Colorbox, Superfish) that I’m not at all familiar with. Have you seen any sites like this before? Any suggestion for what I might be able to exclude without losing content?”

Operationalizing Relevance

In traditional IR, the relevance of a document to a specific query is determined through the following process:

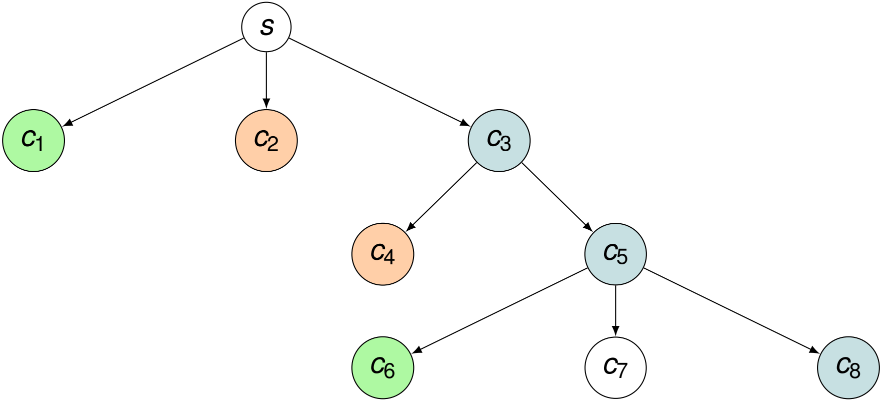
1. Research subjects are presented with a query. This is often a question or topic someone would like to know more about
2. Subjects are presented with a number of documents that might be related to the query. They judge each document as being “relevant” or “not relevant” to the topic. This human judgment is regarded as the “ground truth”
3. The researchers then design an IR system that best approximates human judgments of relevance

Operationalizing Relevance II

This process is not applicable to the web archives for a multitude of reasons.

- AIT interface, which clients use to create and manage their own web archives, does not have a query interface that can be used to execute topic-based queries.
- AIT clients do not query their own web archives in the traditional sense; they judge whether an archived webpage or website is relevant or not by looking at crawl reports
- In general, web archives rarely have full-text indexing

Operationalizing Relevance: Depiction of a Website as a Link Graph



Operationalizing Topic Relevance, Definition 1

$$TR(c_i) = \frac{1}{d(s, c_i)}$$

$$TR(c_3) = \frac{1}{d(s, c_3)} = \frac{1}{1} = 1$$

$$TR(c_4) = \frac{1}{d(s, c_7)} = \frac{1}{3} = 0.333$$

Higher values indicate higher relevance

Operationalizing Topic Relevance, Definition 2

The seed domain is represented as a vector *S*. Other domains that are linked to from the seed domain are represented as vectors *D*. The topic relevance is the cosine similarity between *S* and *D*

$$cosine(S, D) = \frac{S \cdot D}{\parallel S \parallel \parallel D \parallel}$$

Example

Seed domain is www.loc.gov, which links to domain cdn.loc.gov
Seed domain is represented as vector
$S =< 1, 0, 0, 1, 1, 0, 2, 0, 1, 3, 2 >$, while
$D =< 1, 1, 0, 1, 1, 1, 2, 0, 0, 1, 0, 2 >$
$cosine(S, D) = 0.30$

Operationalizing Size Relevance

Archived website can be larger than the original website, but not *much* larger. The difference in cardinality between $|A|$ and $|O|$ must not exceed a certain user-defined limit, defined as k

$$|A| - |O| \leq k$$

Each web archivist determines how much larger she thinks the archived website can be when compared to the original

Archivability

- The intrinsic properties of a website that make it more difficult to archive
- In the data, archivability was a dimension of IQ that was perceived by AIT employees (that is, web archivists), but rarely by AIT clients
- Web archivists need to know if a website is archivable *before* capturing it in order to ensure a high-quality archived website
- Archivability is thus a *latent* dimension, because it is hidden from most people

Examples of Problems with Archivability

Example 1

"For Facebook, your site was archived, there is just an issue that is keeping the archived page from displaying normally. Our engineers are working on this and it should be fixed this week"

Example 2

"It looks like the site uses a fair bit of javascript to generate those 'printer friendly' pages, but I'm not sure how feasible capture is"

Example 3

"Because of their interactive nature, search boxes cannot operate in an archived website in the same way as they would on the live web"

Operationalizing Archivability

Archivability of a website is defined as a measure of *dynamism* D_O, or the number of dynamic components contained in the original website. The higher the dynamism of a website, the lower its archivability.

$$Archivability = \frac{1}{Dynamism} = \frac{1}{|D_O|}$$

Example

A website has nine components, therefore $|O| = 9$
Two of these components are JavaScript files, and so are dynamic; $|D_O| = 2$
Archivability = $2/9$ or 0.22

Unexpected or Surprising Findings

- In the final theory, completeness is a sub-dimension of IQ (part of the correspondence dimension) rather than having its own dimension
- AIT clients often made mistakes when judging whether or not archived content was relevant, regularly flagging web content as irrelevant, when it was actually relevant
- AIT clients were worried as much about the overabundance of content in their web archives as about their completeness. This directly contradicted the researcher's initial assumption that the size of a web archive would not be important for them
- Archivability was a latent dimension. Only AIT employees, not clients, were able to determine a website's archivability and judge how it will affect the quality of its archived counterpart

Applying the Operationalized Definitions of IQ

1. Develop software that can apply the IQ metrics
2. Carry out experiments to determine which metrics perform best. Ex: cosine or Jaccard similarity for completeness
3. Refine the metrics. Correct formula for an IQ dimension might be more complex than originally stated. Ex: Visual correspondence was defined as $VC(O, A) = \frac{1}{ed(O, A)}$. In reality, the formula might be something closer to $VC(O, A) = \frac{\alpha}{ed(O, A)}$

The most difficult challenge lies, not in applying the metrics, but in securing access to a dataset that would allow these metrics to be used and refined

Other Research Directions

- Investigate in depth the mismatch between how humans perceive web archives and how web archives are usually constructed
- Explore the notion of time and its effect on the IQ of a web archive

Conclusion

Thank you for your time and support.

References I

[1] Ainsworth, S. G., & Nelson, M. L. (2015). Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. *International Journal on Digital Libraries, 16* (2), 129?144. Retrieved from http://dx.doi.org/10.1007/s00799-014-0120-4 doi: 10.1007/s00799-014-0120-4

[2] Ainsworth, S., Nelson, M. L., & Van de Sompel, H. (2014). A framework for evaluation of composite memento temporal coherence. *Computing Research Respository (CoRR), abs/1402.0928*. Retrieved from http://arxiv.org/abs/1402.0928

References II

[3] AlNoamany, Y., Weigle, M. C., & Nelson, M. L. (2015). Detecting off-topic pages in web archives. In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Research and Advanced Technology for Digital Libraries: Lecture Notes in Computer Science* (Vol. 9316, pp. 225-237). Cham, Switzerland: Springer International Publishing. Retrieved from: http://dx.doi.org/10.1007/978-3-319-24592-8 17 doi: 10.1007/978-3-319-24592-8_17

[4] Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2016). *Web archiving in the United States: A 2016 Survey* (Research Report). Retrieved from http://ndsa.org/publications

References III

[5] Banos, V., & Manolopoulos, Y. (2015). A quantitative approach to evaluate website archivability using the CLEAR+ method. *International Journal on Digital Libraries*, 1-23. Retrieved from http://dx.doi.org/10.1007/s00799-015-0144-4 doi:10.1007/s00799-015-0144-4

[6] Brunelle, J. F., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, 1-19. Retrieved from http://dx.doi.org/10.1007/s00799-015-0150-6 doi: 10.1007/s00799-015-0150-6

References IV

[7] Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011, March). The SHARC framework for data quality in web archiving. *The VLDB Journal, 20* (2), 183?207. doi:10.1007/s00778-011-0219-9

[8] Glaser, B., & Strauss, A. (1967) *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick, NJ: Aldine Transaction.

[9] Kuny, T. (1997, September). *A digital dark ages? Challenges in the preservation of electronic information*. Presented at the 63RD International Federation of Library Associations and Institutions (IFLA) Council and General Conference, Copenhagen, Denmark. Retrieved from http://archive.ifla.org/IV/ifla63/63kuny1.pdf

References V

[10] Lazarsfeld, P.F. (1959). *Problems in methodology*. In R. Merton, L. Broom & L. Cottrell (Eds.), *Sociology today* (pp. 39-78). New York: Basic Books.

[11] Masanés, J. (2006). *Web archiving*. Berlin; New York: Springer.

[12] Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web.*Journal of the American Society for Information Science and Technology, 53*(2), 145-161. doi: 10.1002/asi.10017

References VI

[13] Spaniol, M., Denev, D., Mazeika, A., Weikum, G., & Senellart, P. (2009). Data quality in web archiving. In *Proceedings of the 3rd workshop on information credibility on the web* (pp. 19?26). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1526993.1526999 doi: 10.1145/1526993.1526999

[14] Stvilia, B. (2006). *Measuring information quality* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

[15] Swiss National Library. (2015). *Visual quality indicator (VQI)* (Tech. Rep.). Bern, Switzerland.

[16] Taylor, R. S. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing Corporation.

References VII

[17] Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.288-295). doi:10.1145/345508.345602