The University Of Sheffield.

Understanding assessors: a forgotten element of big data research?

Robert Villa
Information Retrieval Group
Information School
University of Sheffield

Big data

- Various definitions
  - Ward and Barker (2013), “Undefined By Data: A Survey of Big Data Definitions”, http://arxiv.org/abs/1309.5821
  - “3V” (Volume, Velocity, Variety) definition is popular

    Douglas, L. (2001). 3d data management: Controlling data volume, velocity and variety. Gartner.
- In many definitions of “big data”, machine learning is core

Machine learning & Training

- Machine learning requires training data
- Often, at some point, this requires humans to annotate or assess data in some way
  - E.g. the many and varied types of annotations used in corpus linguistics, annotations of video shots, image annotations, relevance judgements of documents, etc.

The University Of Sheffield.

Here: focus on “assessors”

Topics / search tasks

Documents

Relevant / Not relevant?

From: https://www.flickr.com/photos/netsrac/173500383

Focus

- Focus is normally on the really cool new technique which is being presented
  - As is normal and would be expected
- Quality of annotations or assessments (e.g. relevance judgements)
  - Inter-rater reliability
- Relatively little work on the assessors / annotators themselves

What we want

Topic	*Iteration*	*Doc No.*	*Relevance*
1	0	AP880212-0161	0
1	0	AP880216-0139	1
1	0	AP880216-0169	0
1	0	AP880217-0026	0
1	0	AP880217-0030	0
...			

From http://trec.nist.gov/data/qrels_eng/

What we want

Topic	*Iteration*	*Doc No.*	*Relevance*
1	0	AP880212-0161	0
1	0	AP880216-0139	1
1	0	AP880216-0169	0
1	0	AP880217-0026	0
1	0	AP880217-0030	0
...			

From http://trec.nist.gov/data/qrels_eng/

Information Retrieval Test Collection

- Composed of:
  - Document collection
  - List of topics
  - List of relevance judgements (“qrels”)

The University Of Sheffield.

We need the help of …

From: https://www.flickr.com/photos/netsrac/173500383

The University Of Sheffield.

Or a crowd if crowdsourcing

From: https://www.flickr.com/photos/sundve/3744159600

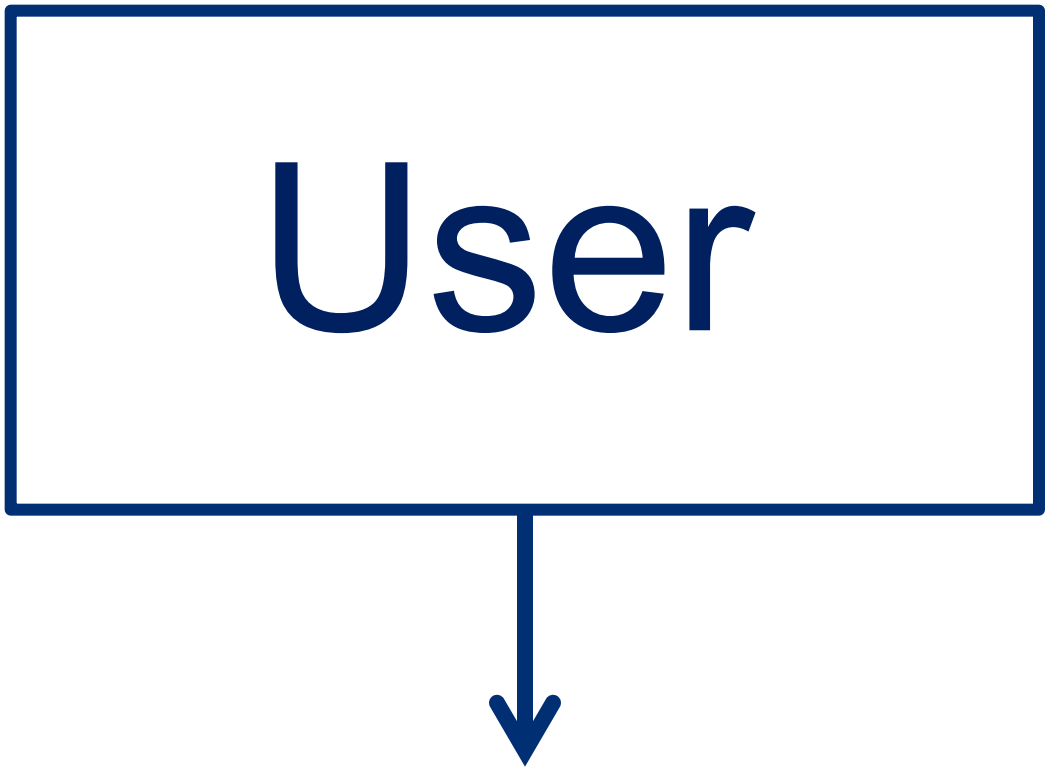
The University Of Sheffield.

Although a crowd is …

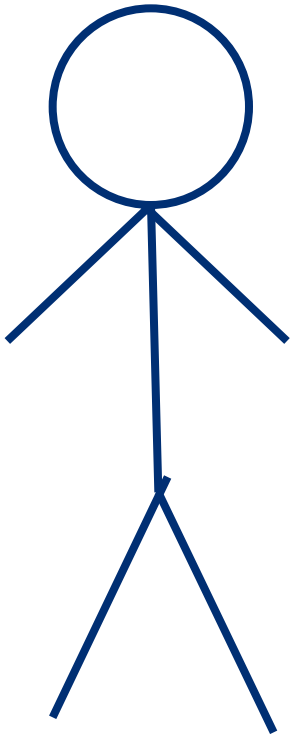
(hopefully)

The University Of Sheffield.

Written up in papers as



Or maybe



User

What exactly are assessors or annotators doing?

- Or who are they?
  - Same people who wrote the paper?

The University Of Sheffield.

The importance of annotators?

- Given the importance of annotations/ assessments for training and evaluation, shouldn't we consider the assessor in more detail?
  - Especially where the judgements being gathered have relatively low inter-rater reliability, e.g. document relevance.

Work of Smucker et. al. at Waterloo

- Aiman L. Al-Harbi and Mark D. Smucker, "User Expressions of Relevance Judgment Certainty", HCIR 2013.
- Mark D. Smucker and Chandra Prakash Jethani. "Time to Judge Relevance as an Indicator of Assessor Error," In the Proceedings of the 35th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, August 2012, Portland, OR, USA. 2 pages.
- Mark D. Smucker, Xiaoyu Sunny Guo, and Andrew Toulis, "Mouse Movement During Relevance Judging: Implications for Determining User Attention," SIGIR 2014, 4 pages.
- C. Jethani and M. D. Smucker. Modeling the time to judge document relevance. In Proceedings of the SIGIR'10 Workshop on the Simulation of Interaction, 2010.

The University Of Sheffield.

Assessor effort – text documents

- Relevance judgements, TREC HARD
- Does the size and relevance of a document affect the degree of effort?
  - Document length was found to affect the effort required to judge a document, but does not affect the accuracy
  - Degree of relevance of a document was found to affect both accuracy and effort: the trend is for accuracy to decrease for relevant documents, and perceived effort to increase.

Robert Villa and Martin Halvey. 2013. Is relevance hard work?: evaluating the effort of making relevant assessments. SIGIR '13. ACM, 765-768.

The University Of Sheffield.

Understanding the annotation process: annotation for Big data

- AHRC funded project
  - Co-PI Martin Halvey, University of Strathclyde
- Can we:
  - Understand more about assessor behaviour?
  - Reduce the effort in constructing assessments?
  - Construct “augmented test collections”?

The University Of Sheffield.

Understanding the assessment process

- Exploration: start with small-scale, qualitative lab-based studies of assessors
  - Own information needs
  - Secondary topics (TREC)
- Medium scale collection of assessments
  - In lab, but larger scale, less detail
- Larger scale collection of assessments
  - Crowdsourcing

The University Of Sheffield.

Initial small scale study: Research Questions

1. What differences are there in primary and secondary relevance judgements?
2. Does the type of search task (open v.s closed) influence relevance?
3. How does topic knowledge, task interest and/or judgement confidence relate to relevance assessment?

The University Of Sheffield.

Gathered search tasks

- Asked participants to provide to recent information needs, one open and one closed
  - 1 sentence description of task
  - More detailed description e.g. context
  - Describe what would constitute a relevant or irrelevant result
  - Some example search terms

Document Collection

- Keywords from task generation used to search Google
- 3 random results were sampled from the first 10 pages of search results, giving 30 documents
- Documents processed using Readability API to remove adverts and other superfluous information
- 15 of the initial documents were chosen for use in the evaluation

https://www.readability.com/developers/api

The University Of Sheffield.

System Interface

Topic 14-2

Description: What led to the recession that began in 2008?

Situation: The economic recession that suddenly occurred throughout the world in 2008 made little sense to anybody outside of economics/finance. the responsibility is still debated and how the effects spread from one country to the next so quickly is also hard to understand. I wanted to get a better understanding of how such an event can occur i.e. what features of current economics/finance allowed such a problem to happen.

Criteria: A relevant document or website would include information about the recession, which countries experienced the economic downturn and what principles of current economics allowed the propagation of the problem throughout the world. Pages that include only information about the period during which the recession took place are not relevant.

How much do you know about this topic?

I know nothing about this topic ○ ○ ○ ○ ○ ○ ○

View document

Topic: Topic 14-2

<< Click to view topic

Document No. 1

Financial crisis - Mises Wiki, the global repository of classical-liberal thought

The term **financial crisis** refers to a variety of situations in which some financial institutions or assets suddenly lose a larg part of their value.*[citation needed]*

They include sovereign defaults, which occur when a government fails to meet payments on its external or domestic debt obligations or both. Then there are banking crises, typically when a significant part of a banking sector has become insolvent after heavy investment losses, banking panics, or both. Another important class of crises consists of exchange rate crises, where the value of a country's currency falls precipitously, often despite a government "guarantee" that it will not allow this to happen under any circumstances. Some crises are marked by bouts of very high inflation. These separate types of crisis often occur in clusters.[1]

Properties[edit]

Carmen Reinhart and Kenneth Rogoff argue, that systemic banking crises are typically preceded by asset price bubbles, large capital inflows and credit booms.

Is this document relevant to the topic?

Not Relevant ○ Relevant ○

How confident are you in making this judgement?

Not confident at all ○ ○ ○ ○ ○ ○ ○ Very confident

Have you seen this document before?

No ○ Maybe ○ Yes ○

Continue >>

The University Of Sheffield.

Study Measures

- Per document
  - Relevance
  - Confidence
  - Seen before
  - Time
  - View topic
- Per topic
  - Knowledge
  - Interest
  - NASA-TLX

Study Procedures

- Each participant judged 15 documents for 6 topics, total of 90 judgements
- No time limits given
- Session recorded with Morae
- Post session Morae was used to play back parts of the session where participants described what they were doing
- Exit interview was conducted during play back

Morae: www.techsmith.com/morae.html

Study Participants

- 20 participants
- Average age ~28 (std dev=8.17, min=19, max=54)
- 9 male, 11 female
- 12 native English speakers, 5 fluent, 3 intermediate
- 14 high search experience, 6 intermediate

Relevance Judgements

- 1200 secondary, 600 primary
- 240 documents had more than 1 assessment (1 primary, 5 secondary)
- 79.5% of secondary agree with primary
- Preliminary results: data still being analysed

The University Of Sheffield.

Primary vs. Secondary

	Primary		Secondary	
	Median	Mean (SD)	Median	Mean (SD)
Relevant	**0**	**0.448 (0.498)**	**0**	**0.45 (0.497)**
Confidence	**7**	**6.302 (1.036)**	**6**	**5.71 (1.4)**
Time (secs)	23.7s	34.6s (34.0s)	26.5s	38.7s (40.3)
Knowledge	**2**	**5.25 (1.391)**	**1**	**2.063 (1.47)**
Interest	**6**	**5.475 (1.569)**	**4**	**3.513 (1.646)**

The University Of Sheffield.

Open vs. Closed

	Closed		Open	
	Median	Mean (SD)	Median	Mean (SD)
Relevant	**0**	**0.370 (0.486)**	**1**	**0.527 (0.5)**
Confidence	**7**	**6.05 (1.244)**	**6**	**5.77 (1.377)**
Time (secs)	**22s**	**30.4s (30s)**	**31100**	**44.2s (44.0s)**
Knowledge	2.5	3.117 (2.148)	2	3.133 (2.038)
Interest	**4**	**3.933 (1.716)**	**5**	**4.4 (1985)**

The University Of Sheffield.

Confidence in Judgement

- Participants tended to report having high confidence in judgements
  - Both of their own topics and other people's
- Although in the qualitative interview it appeared that assessors interpreted "confidence" in different ways

The University Of Sheffield.

RQ1: Differences in Primary and Secondary Assessors

- Some disagreement between primary and secondary assessments
- Also difference in secondary judgements disagreeing with “gold standard”
- Key theme from interviews was difficulty in interpreting context and scope of open tasks

RQ2: Differences based on task type

- Open vs Closed is only one of many representations
- Significant differences for a number of measures
- TLX indicated that open were more mentally demanding
- Task type needs to be considered when building test collections and also when evaluating assessors

The University Of Sheffield.

RQ3: Knowledge, interest or confidence impact judgement?

- Confidence level had no impact
- Knowledge led to faster judgements
- High interest level led to longer judgement times
- Interviews revealed that assessors with high interest began reading articles for personal reasons

The University Of Sheffield.

Assessor effort – images

- ImageCLEF 2007 collection
- Does image size, topic difficulty and semantic/visual nature of the topic affect assessor effort?
  - Image size affects the time required to judge an image (larger images → more time) but does not affect accuracy or perceived effort
  - Topic difficulty affects accuracy, time, and effort: trend for accuracy to decrease as difficulty increases, time and perceived effort to increase as difficulty increases
  - As topics move from being visual to semantic, accuracy decreases, time to judge and perceived effort increases

Martin Halvey and Robert Villa. 2014. Evaluating the effort involved in relevance assessments for images. SIGIR '14. ACM, 887-890.

Tracking assessors

- In all of these studies, we have attempted to log assessors
- If this data is valuable, could it become part of the test collection?

Augmented test collections

- Can we, or should we, augment conventional test collections with extra data about the assessors?
  - E.g. log assessor actions
  - “Big data” often involves the processing of user activities (search logs, consumer buying behaviour, books bought, etc.)

The University Of Sheffield.

For IR the proposal is to include:

- Information about the assessors
  E.g. demographics, etc.
- The relationship between the assessor and topic
  E.g. topic expertise, etc.
- How the assessors went about the judging of the documents
  E.g. confidence in the assessment, time to assess, etc.
- Represent multiple points of view
  E.g. keep judgements from multiple assessors

Halvey, M., Villa, R. and Clough, P. (2014) SIGIR 2014 workshop on gathering efficient assessments of relevance (GEAR), SIGIR '14, Gold Coast , Australia, July 2014, 1293-1293.

But why?

- Active learning
  - Machine learning system chooses which instances are used for training
  - E.g. Settles, B., Craven, M. & Friedland, L. (2008). Active learning with real annotation costs. Proceedings of the NIPS Workshop on Cost-Sensitive Learning, 1069-1078.
- Delays decisions: potentially allows greater flexibility when evaluating
  - Make the subjectivity and complexity of the relevance judgement process more explicit. E.g. do multiple assessors disagree?
- If the assessor is the “ruthless abstraction” of a user, create a slightly less ruthless abstraction
  - The assessor is playing the role of the user in TREC style evaluation. Search engines track users, why not track assessors?

The University Of Sheffield.

Summing up

- The importance of assessors are sometimes over looked
- Can sssessor behaviour and background be important?
- Can we augment collections with more information about how the data was created?

The University Of Sheffield.

Acknowledgements

- Dr. Martin Halvey, Dr. Simon Wakeling, and Laura Hasler
- This work was funded by the UK Arts and Humanities Research Council (grant AH/L010364/1)

The
University
Of
Sheffield.

The University Of Sheffield.

Annotators / Assessors

- Focus on annotations rather than annotators
  - For obvious reasons
- Yes, given the importance of annotators or assessors, should be not be considering their needs in greater detail?

The University Of Sheffield.

AHRC Project: Understanding the annotation process

- Exploration: start with small-scale, qualitative lab-based studies of assessors
  - Own information needs
  - Secondary topics (TREC)
- Medium scale collection of assessments
  - In lab, but larger scale, less detail
- Larger scale collection of assessments
  - Crowdsourcing etc.

The University Of Sheffield.

Conclusions

- A number of benefits to current test collection set up
- It would be possible and beneficial to store/share information about assessors
- Interest and knowledge in particular seem to impact bevaviour
- Can perhaps augment collections with secondary assessors similar to gold standard

Evaluation(s)

- To date
  - Looking at capturing effort for differing documents
  - Looking at capturing effort for differing tasks
- Most recent evaluation
  - Measuring effort for both primary and secondary search tasks
  - Looking at different measures for task “interest”

The University Of Sheffield.

Gold Standard

	Gold	Agree	Disagree
Relevant	0.408 (0) σ=0.493	0.409 (0) σ=0.492	0.594 (1) σ=0.492
Confidence	6.56 (7) σ=0.913	5.860 (6) σ=1.352	5.138 (5) σ=1.436
Time (secs)	41540 (25831) σ=40488.371	36445 (24284) σ=37282.396	50677 (38390) σ=48410.079
View Topic	0.050 (0) σ=0.2018	0.122 (0) σ=0.369	0.187 (0) σ=0.449
Knowledge	4.938 (5) σ=1.393	2.073 (1) σ=1.476	2.020 (1) σ=1.407
Interest	5.687 (6) σ=1.046	3.515 (4) σ=1.599	3.504 (4) σ=1.777

The University Of Sheffield.

Task Knowledge

Know	Count	Relevant	Confidence	Time	V. Topic
1	629	0.432 (0) σ=0.496	5.827 (6) σ=1.436	3 7 6 9 9 . 0 1 3 (26750) σ=39273.924	0.183 (0) σ=0.465
2	285	0.428 (0) σ=0.496	5.607 (6) σ=1.278	5 8 0 3 6 . 8 0 4 (37475) σ=55098.984	0.25 (0) σ=0.591
3	135	0.533 (1) σ=0.501	5.482 (5) σ=1.209	2 7 7 6 7 . 8 1 5 (23244) σ=19105.880	0.193 (0) σ=0.465
4	210	0.419 (0) σ=0.595	5.991 (6) σ=1.237	3 0 8 2 6 . 5 0 5 (21454.5) σ=32906.843	0.152 (0) σ=0.444
5	165	0.467 (0) σ=0.501	5.952 (6) σ=1.521	2 6 3 4 7 . 1 2 7 (19797) σ=21139.056	0.109 (0) σ=0.35
6	270	0.507 (0) σ=0.501	6.267 (7) σ=0.985	3 4 5 7 3 . 3 4 4 (25818) σ=28256.887	0.126 (0) σ=0.374
7	105	.343 (0) σ=0.477	6.619 (7) σ=.0.826	2 8 9 8 0 . 9 2 4 (20111) σ=26861.376	0.095 (0) σ=0.326

The University Of Sheffield.

Task Interest

Interest	Count	Relevant	Confidence	Time	V. Topic
1	195	0.43 (0) σ=0.496	5.59 (6) σ=1.655	3 7 2 8 9 . 9 3 (22370) σ=48242.75	0.06 (0) σ=0.262
2	225	0.427 (0) σ=0.496	5.782 (6) σ=1.477	4 3 3 6 0 . 6 1 8 (31135) σ=40377.141	0.161 (0) σ=0.412
3	240	0.404 (0) σ=0.492	5.782 (6) σ=1.279	3 4 6 5 5 . 5 7 5 (23201) σ=34498.388	0.094 (0) σ=0.362
4	270	0.470 (0) σ=0.500	5.711 (6) σ=1.225	3 1 8 5 6 . 6 1 4 (23372) σ=30441.278	0.039 (0) σ=0.194
5	345	0.420 (0) σ=0.494	6.133 (7) σ=1.169	3 4 7 1 5 . 5 8 6 (23763) σ=38577.813	0.061 (0) σ=0.240
6	345	0.461 (0) σ=0.499	5.971 (6) σ=1.232	4 1 1 4 7 . 6 5 5 (27604) σ=29587.420	0.028 (0) σ=0.165
7	180	0.539 (1) σ=0.500	6.322 (7) σ=1.156	3 9 6 1 7 . 5 8 3 (30412.5) σ=29587.420	0.061 (0) σ=0.240

The University Of Sheffield.

Understand more about assessor behaviour

- Effort: how much effort do assessors put in to making judgements
- Assessor
  - What do assessors do?
- Contextual factors, e.g. background of the assessor
- Confidence in judgements
  - Phycology scales are multiple