Winter 10-17-2020

Content Modelling for unbiased Information Analysis

MILIND GAYAKWAD
Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune,
mdgayakwad@bvucoep.edu.in

Suhas Patil Dr
Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, shpatil@bvucoep.edu.in

Content Modelling for unbiased information analysis

Mr. Milind Gayakwad, Dr. Suhas Patil
Bharati Vidyapeeth (Deemed to be University)
College of Engineering, Pune

Abstract

Content is the form through which the information is conveyed as per the requirement of user. A volume of content is huge and expected to grow exponentially hence classification of useful data and not useful data is a very tedious task. Interface between content and user is Search engine. Therefore, the contents are designed considering search engine's perspective. Content designed by the organization, utilizes user's data for promoting their products and services. This is done mostly using inorganic ways utilized to influence the quality measures of a content, this may mislead the information. There is no correct mechanism available to analyse and disseminate the data. The gap between Actual results displayed to the user and results expected by the user can be minimized by introducing the quality check for the parameter to assess the quality of content. This may help to ensure the quality of content and popularity will not be allowed to precede quality of content. Social networking sites will help in doing the user modelling so that the qualitative dissemination of content can be validated.

Introduction

Information is explored in the form of content on the internet. Contents viewed stored and accessed in the form of various types like facts, opinion or irrelevant content also there are various forms of content namely Text, Image, Audio, Video. Social communication and consideration are important as being human, we tend to rely on a society. Quality of information in unbiased format is extremely desired as lot of day today activities, opinion, decisions are made based on content that we read, watch or listen.

One can express the information in the form of content through their experience or suggestions. Users decisions are often influenced by this source provided by the society; hence we look at review before watching the movie or buying the product. Significance of content can be figured out by the examples of giants in the industry involved in the various phases of content modelling. Organization like Wikipedia generates the content with help of domain expert in the respective field, Google search engine, YouTube helps in exploring the content and Google plus, Facebook let us access with society.

Now a days to address the business needs, political benefit, publicity stunts the legacy of the information in the content is compromised. Research in content modelling is important as we heavily rely on this digital world.

Dissemination of irrelevant data implies the usage of the inorganic way to promote the rank of a content. If this is not mitigated then such irrelevant data may gain popularity and superseded the other qualitative content, which is more hazardous. Using the existing structure (available ways to provide the input, process and output) to gather the information about a user in social networking site and popular search engine the modelling of content is done. This research work is targeted to the general audience to get the desired result in the form of content with priority. Even companies dealing with the advertisement can take the advantage as only qualitative content will get appreciation. There is a big challenge to access the content of the website of some other person. Here we may face a problem of restriction so we c start our content generation in the form of audio, Video, text. This problem can be solved by outsourcing the content generation. The existing search engine has their own prioritization mechanism, but it is not sufficient, so the modelling of a content and modelling of the user with the help social networking sites. The data can only be collected after the consent of a user. These modelling techniques help in studying the patterns, once the identity of content is completed, categorization and relevant predictions can be done. Also, the unseen, unidentified patterns can be derived. This type of arrangement could lead to proper content generation and check whether user have accepted.

The existing search engines have their own prioritization mechanism, but it is not sufficient, so the modelling of a content and dissemination of the content with the help social networking sites is important. The data can only be collected after the consent of a user. These modelling techniques help in studying the patterns, once the identity of content is completed, categorization and relevant predictions can be done. Also, the unseen, unidentified patterns can be derived. This type of arrangement could lead to proper content generation and check the acceptance from a user.

Literature Survey:

To perform the literature survey on content modelling total 52 papers are reviewed out of it 5 more relevant to area of interest, from authentic source of publication and recent years are disused. Effect of this type of content would result in misleading a user in taking decision or making opinion about something.[1] Information could be influenced and twisted as per the individual's benefit [2][3]. Edited, partial or presenting the hate speech may result into harming the feelings of addressed community, party or religion.[4][5][6]

Fasten the dissemination of correct, true information and apply the countermeasure for context violating social code of conduct.[7][8]

- **Content analysis based on Foraging Theory-** This mechanism of content modelling helps in identifying intention of the user for [9] by noting the Positive, Negative voting intension analysis using Ant Food Foraging analogy[1]

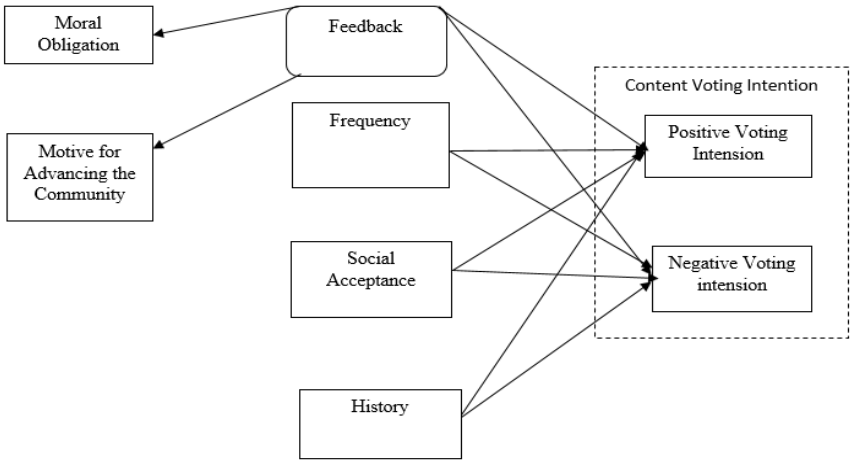

Figure 1: Content analysis based on Foraging Theory

- **Semantic Social Network Analysis by Cross-Domain Tensor Factorization** –Tensor factorization of Users, Tweets, Topic (User-Topic-Vote extraction for individual Tensor Formation) extracts the intersection by plotting the graph like structure of the data. This research proposes the formation of database storing this type of Non-English Information in the form of DBPedia.[2]

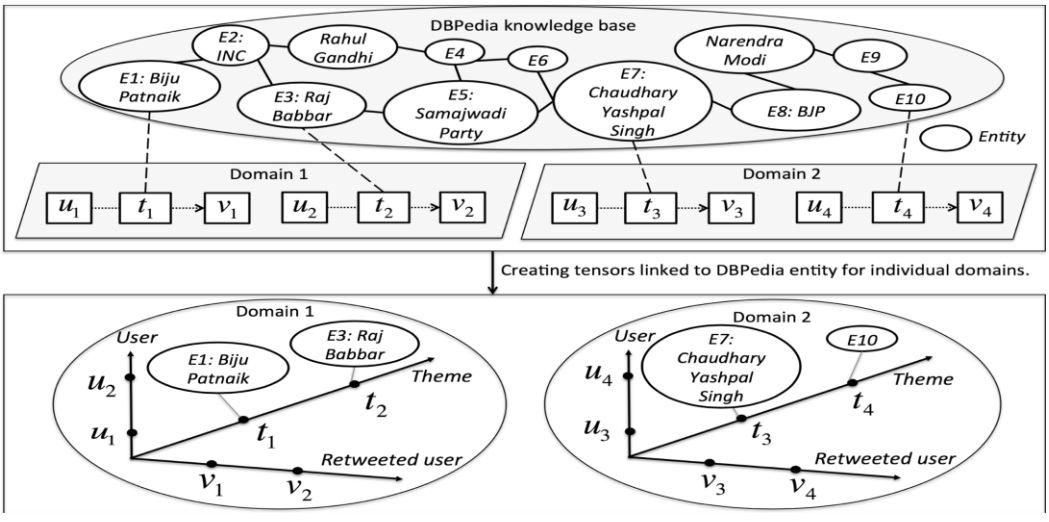

Figure 2: Semantic Social Network Analysis by Cross-Domain Tensor Factorization

- **Collaborative Filtering-Based Recommendation of Online Social Voting-** This is supervised version of recommender system. On a social networking platform votes are collected to decide the recommendation for a topic. This experiment was performed for Tweeter, so tweets and retweets are considered as a matrices for voting.[3]

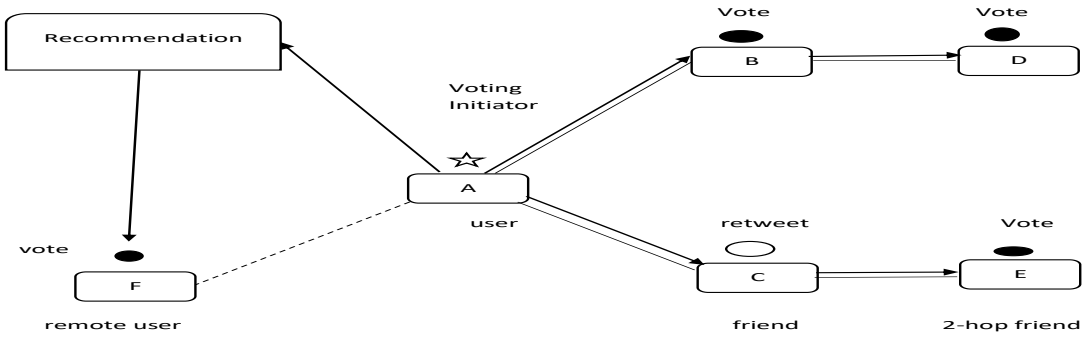

Figure 3: Collaborative Filtering-Based Recommendation of Online Social Voting

- **People, Technologies, and Organizations Interactions in a Social Commerce Era** – The blend of social and technical expertise is considered as a decision-making combination before starting the purchase of product or a service. Human belives on a people to whome he or she knows especially. factor fromTechnical facilitation adds extra clarity in making the decision like analysis and comparison can be only done effectively with the help of relevant technology.[4]

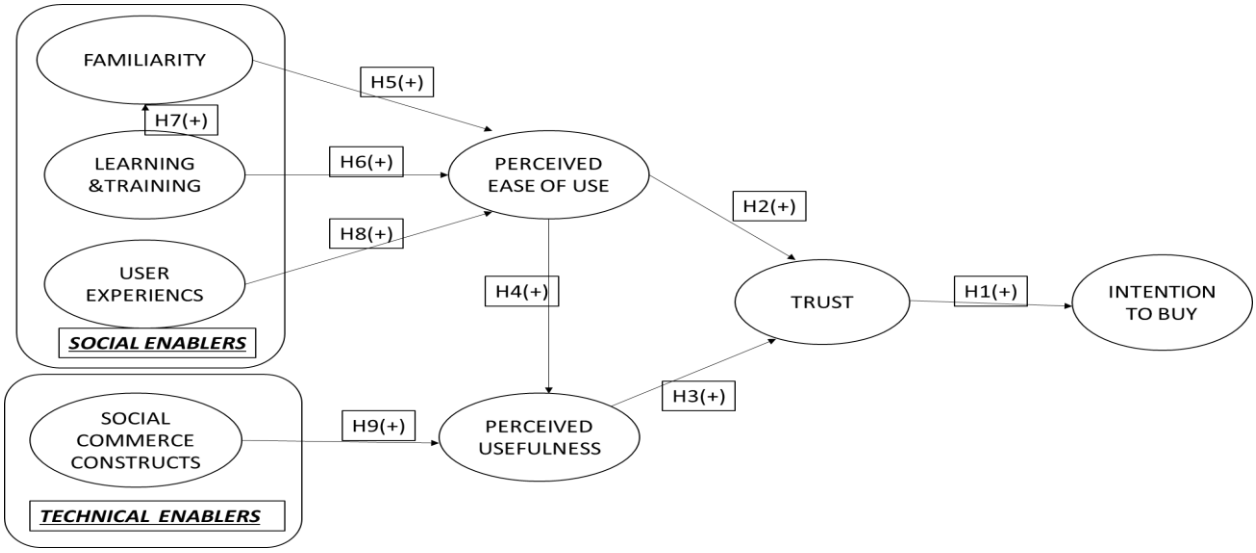

Figure 4: People, Technologies, and Organizations Interactions in a Social Commerce Era

- **Low-rank Multi-view Embedding Learning for Micro-video Popularity Prediction** – Author mentioned various forms of the content

like Visual, Acoustic, Text. These formats have their own mechanism to extract the Social popularity prediction.[5]

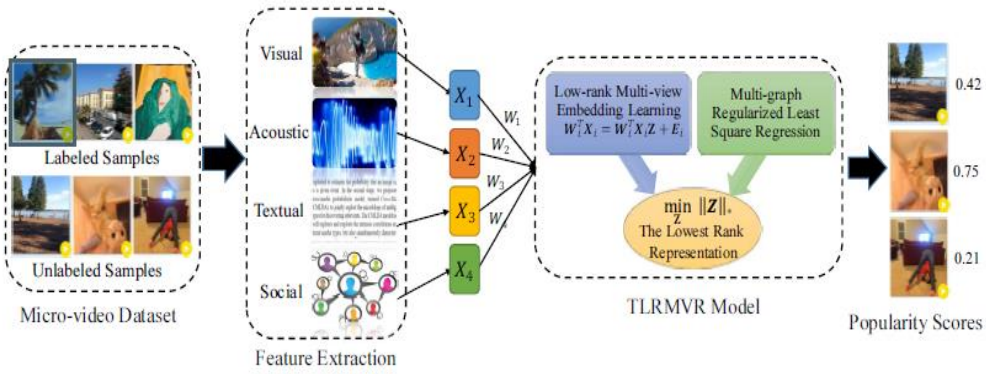

Figure 5: Low-rank Multi-view Embedding Learning for Micro-video Popularity Prediction

Bibliometric Survey

Analysis of Scopus Indexed Journals papers made from year 1974-2020 to understand the scope and research going on content modeling. Various parameters are observed during the analysis –

1) Year wise count of publication
2) Funding received to different organizations
3) Various domains in which the study is carried out
4) Research work across the globe on content modeling
5) Source wise List
6) University or Organization wise List
7) Author wise
8) Keyword Analysis
9) Document Types
10) Year wise Publication in various domains
11) Keyword Mapping

This analysis is useful to know the International status of the topic, current advancement and trend to work on the respective domain.

1) Year wise count of publication

The analysis provides the year wise publication statistics. This helps in analysis of the scope and trend associated with the topics pertaining to the content modeling. 2019 was the year, where maximum number of papers published and in 2020 count is decreased by 8 numbers.

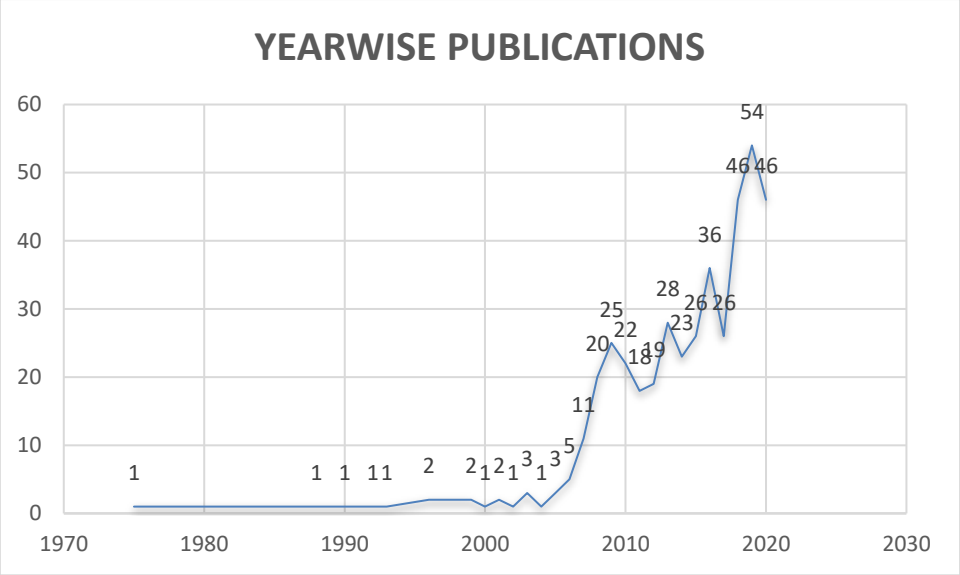

Figure 6: Year wise Publication on documents published on Content Modeling from year 1970-2020

2) Funding received to different organizations

Funding details received for performing the research associated with the content modeling with the agency, which funded the project is mentioned below.
This analysis depicts the relevance of the topic to the industry and government. University of South Florida worked (or working) on 331 research Grants.

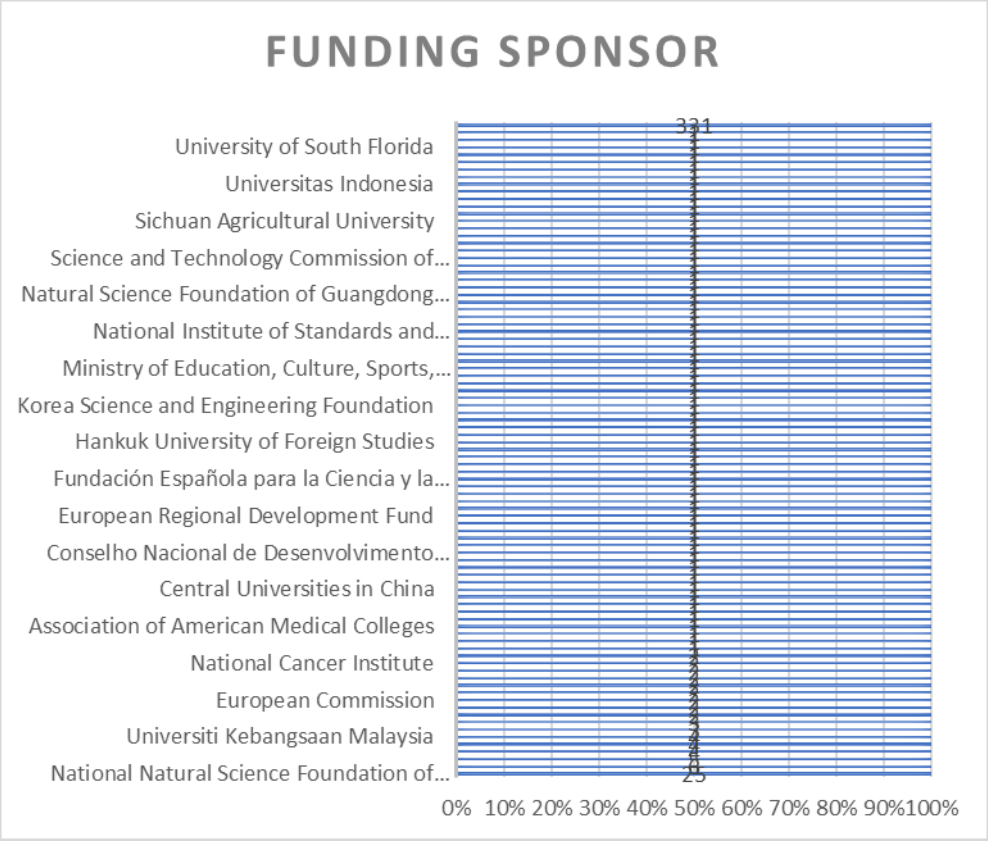

Figure 7: Funding from various organizations.

3) Various domains in which the study is carried out

Applications of the problem statement is highly essential to examine the use of the research to be accomplished to the society.
As given in the pie chart Computer Science, Engineering, Social Sciences are top three areas getting benefitted.

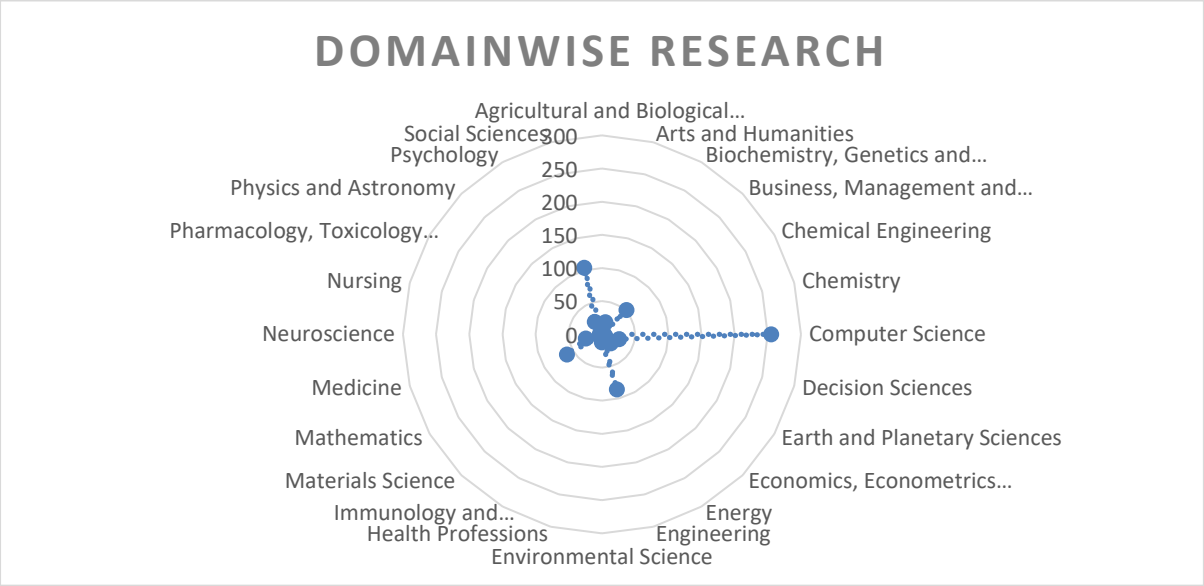

Figure 8: Domain wise publication on Content Modeling

4) Research work across the globe on content modeling

Research undergoing in various countries for the content modeling is given in the "Documents by territory".
This analysis is important in studying the country wise progress in the research associated with content modeling. United States, China and Japan are amongst countries working on maximum number of researches.

Figure 9: Country wise publication

5) Source wise List

To perform the detailed analysis of the Content Modeling, source of the document used for the study is important.
This type of analysis helps in analyzing the research work published at various Journals, Conferences, Notes etc. Study of the timespan 2002 to 2020 total 17 years, is used for the analysis

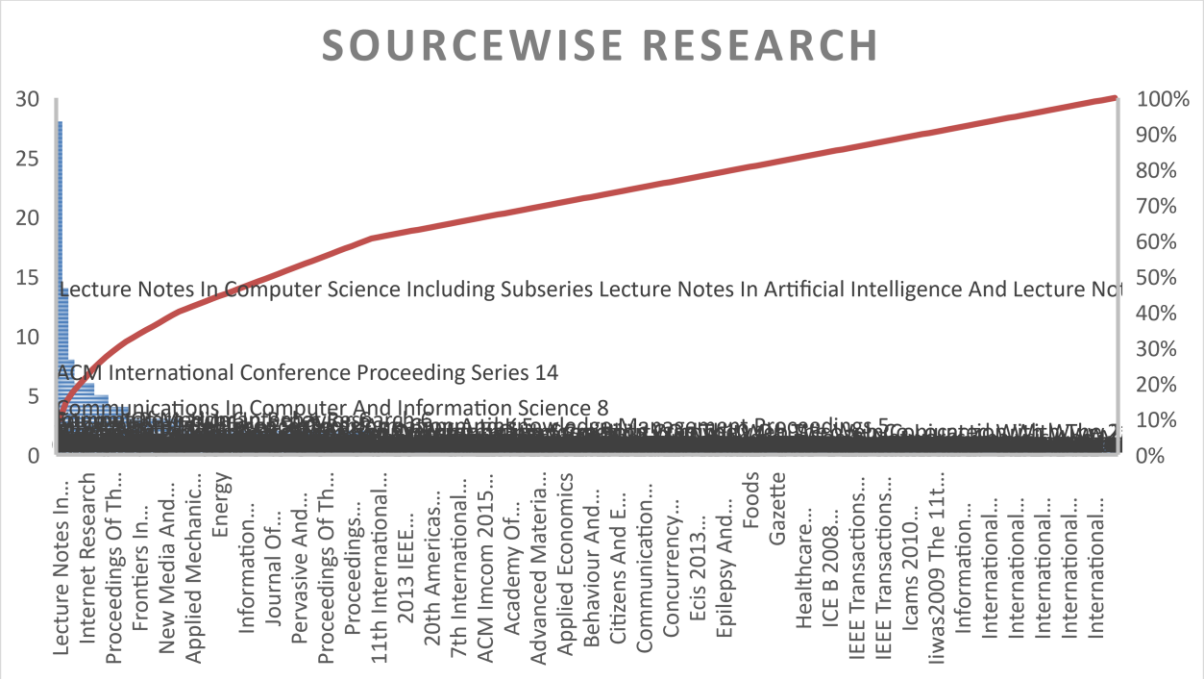

Figure 10: Source wise publication on Content Modeling

6) University or Organization wise List

University contribution in performing the research on Content Modeling is given. Universities with their documents available is given below.
University wise study helps in establishing the association for the collaboration is important.

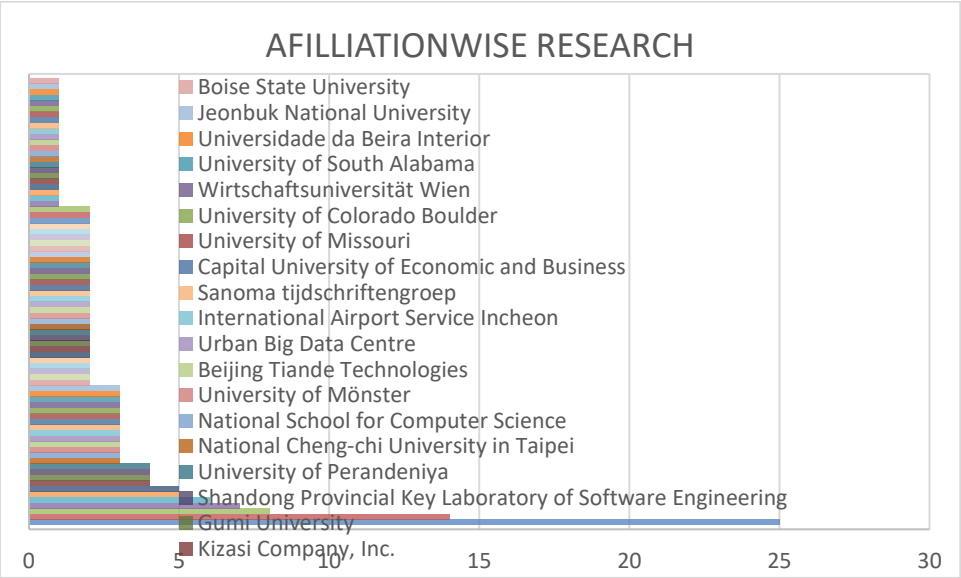

Figure 11: organization wise Publications

7) Author wise List

List of authors performing the research on content modeling is given in the bar graph mentioned below. Authors instrumental in the content modeling is given.

This study helps in examining the benchmarks achieved in the topic of content modeling. Also, one can seek the guideline for the research to be performed.

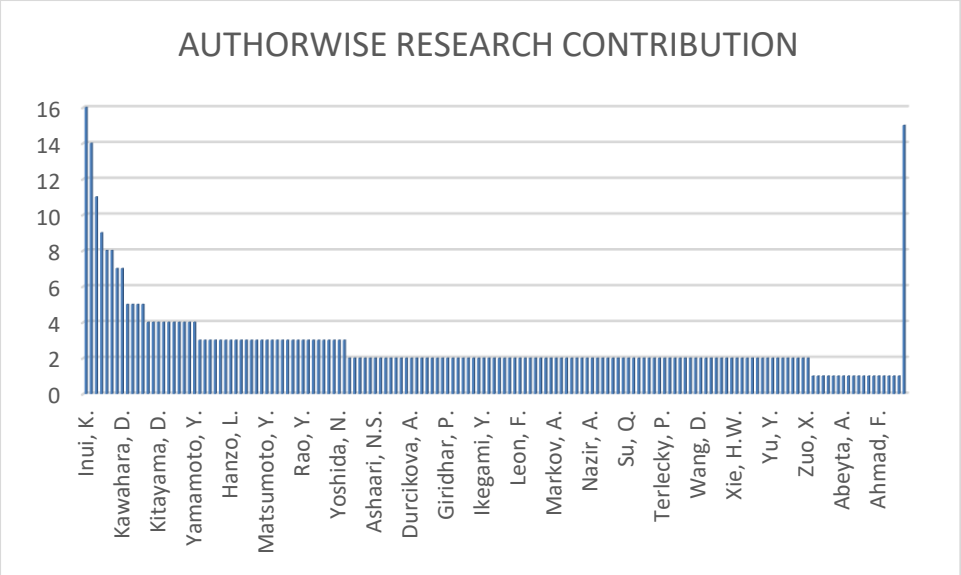

Figure 12: Author wise publications

8) Keyword Analysis

Keywords studied for the analysis of the content modelling are mentioned in this pie chart. Keywords like Information credibility, social networking is covering major chunk of the pie.

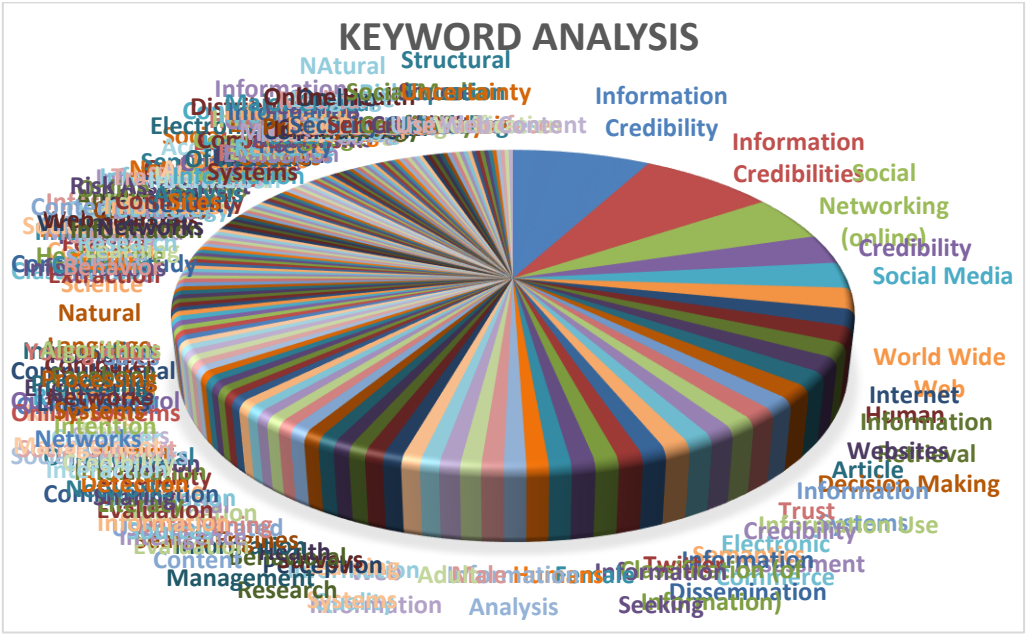

Figure 13: Keyword Analysis

9) Document Types

Various documents from different domain are covered. 223 articles and 168 papers on the conference are covered. Along with this Editorial, Book Chapter, Review, Conference Review, Conference Paper, Article covered.

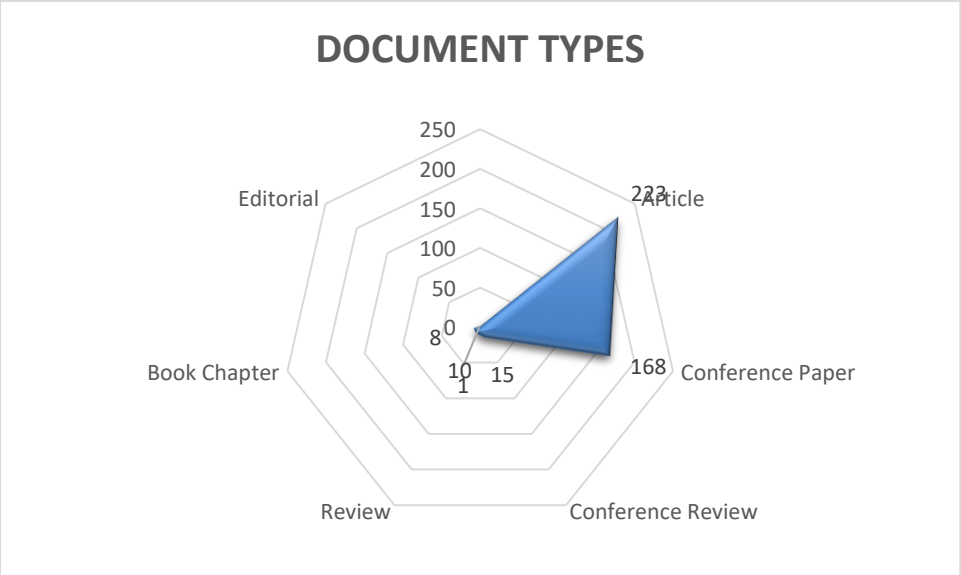

Figure 14: Document Types

10) Year wise Publication in various domains

Contour of year wise publication in various domains is plotted here. Ternary mapping is useful to visualise the per year publication in a domain. Looking at the contour Computer Science is one of the are addressed in a timespan from 2010 to 2020.

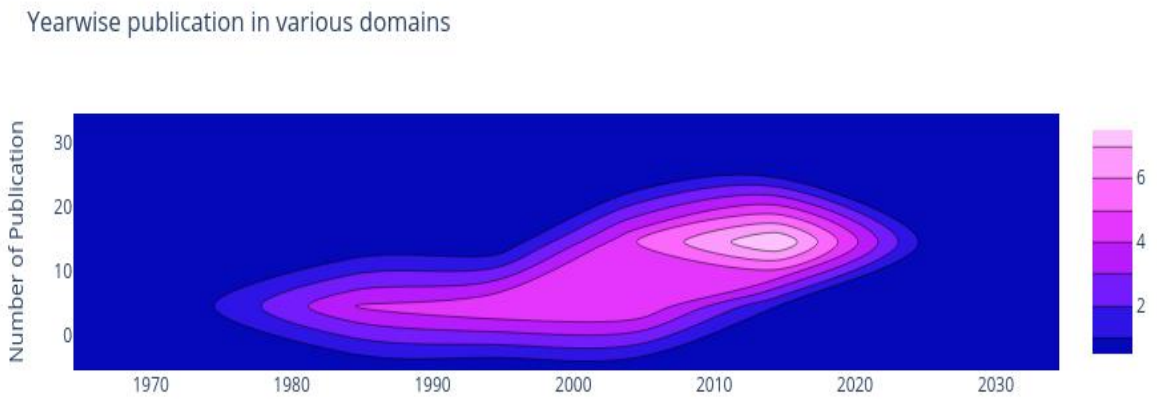

Figure 15: Year wise publications in various domains

11) Keyword Mapping

Mapping of the keyword by using the directed graph helps in studying the association of keywords from this domain. Total 188 nodes and 168 edges are represented here.

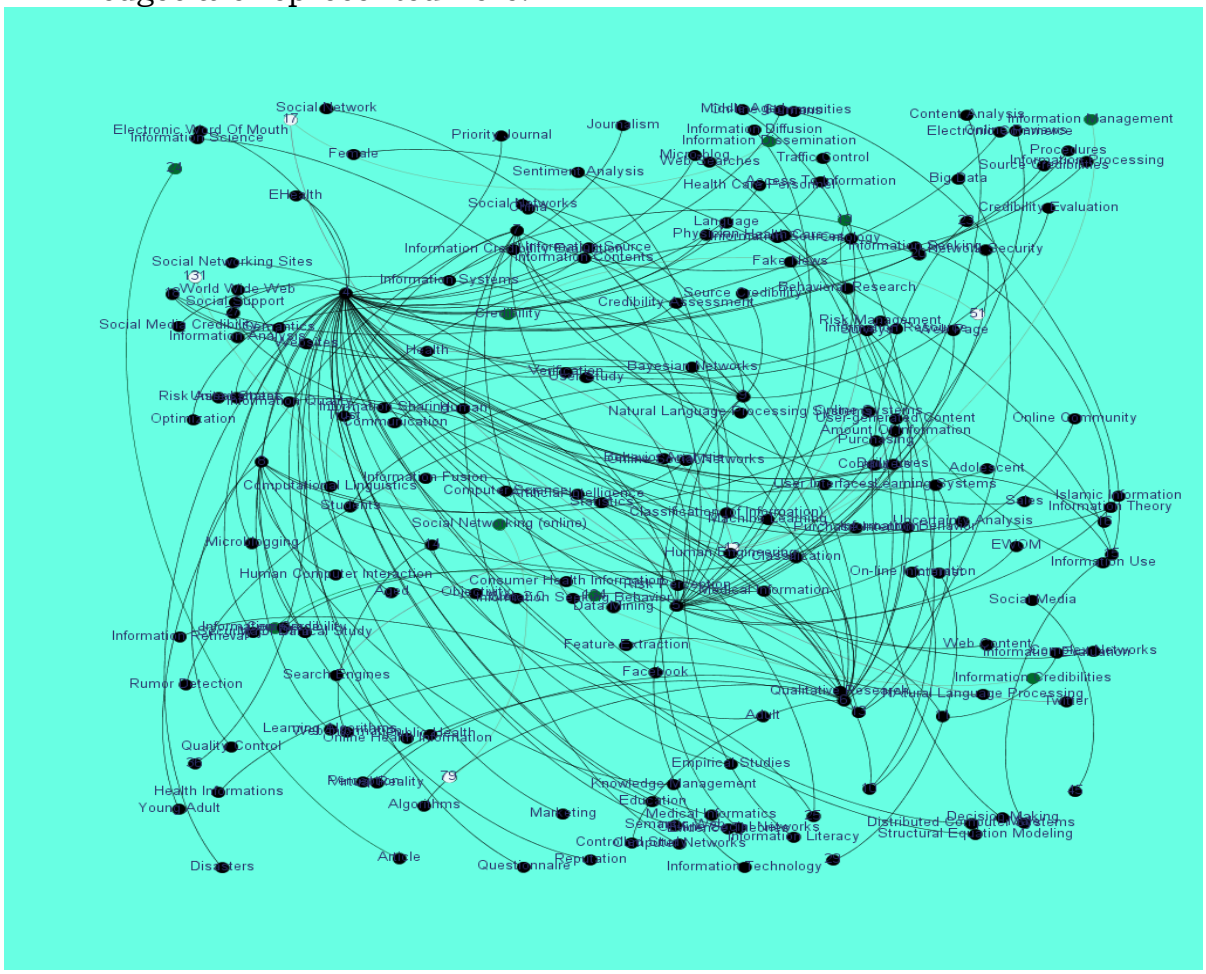

Figure 16: Keyword Mapping

Comparative Analysis of recently published SCI journals

Sr. No.	Research paper	Positive Aspect	Scope for improvement
1	Detecting Fake News Over Online Social Media via Domain Reputations and Content Understanding (2020)	domain reputations (e registration behaviors, registration timing, domain rankings, and domain popularity) and content understanding	TF-IDF and LDA are inefficient in detecting fake news

2	D-AHP method with different credibility of information (December 2017)	e credibility of information in the D-AHP method slightly impacts the ranking of alternatives, but the priority weights of alternatives are influenced in a relatively obvious extent.	measuring the credibility of information will be studied
3	Exploiting Social Review-Enhanced Convolutional Matrix Factorization for Social Recommendation (2019)	rarely consider the user's reviews to capture the user's interests, but in reality, users often express their preferences by posting different reviews to different items	leverage more context features to further improve the recommendation as our work, (context aware system)
4	Information credibility on Twitter (739) citations	Measurements of retweet	Retweet is only applicable for tweeter

Table 1: Comparative Analysis

Research Gap

Available models and approaches focus on the parameters of measuring the quality of information [6][7], Positive or Negative intention of feedback (biased), building trust based on the social and technical [8][9][10]. There is need to focus on dissemination of legit information and its peculation to the mass. Similarly, potentially problematic contents should be prevented to avoid the damage to the society.[11]

Popularity often supersedes Quality of content [12], frequency of doing so in not very clear. That is why the content about nudity [13], hate speech,[14] fake news with some catchy title and thumbnail gets the users attention; which is not good.[15]

Problem Statement

Content is the prominent interface between the Information and web user. Web user's decision is influenced by content; precise content leads to dissemination of Information.

Objectives:

- Data Extraction and collection using standard dataset like content credibility corpus (C3)
- Identify parameters affecting the quality of content
- Model the content and calculate the score
- Validation of content through the user's acceptance to the content

Research Methodology

System Design Outline

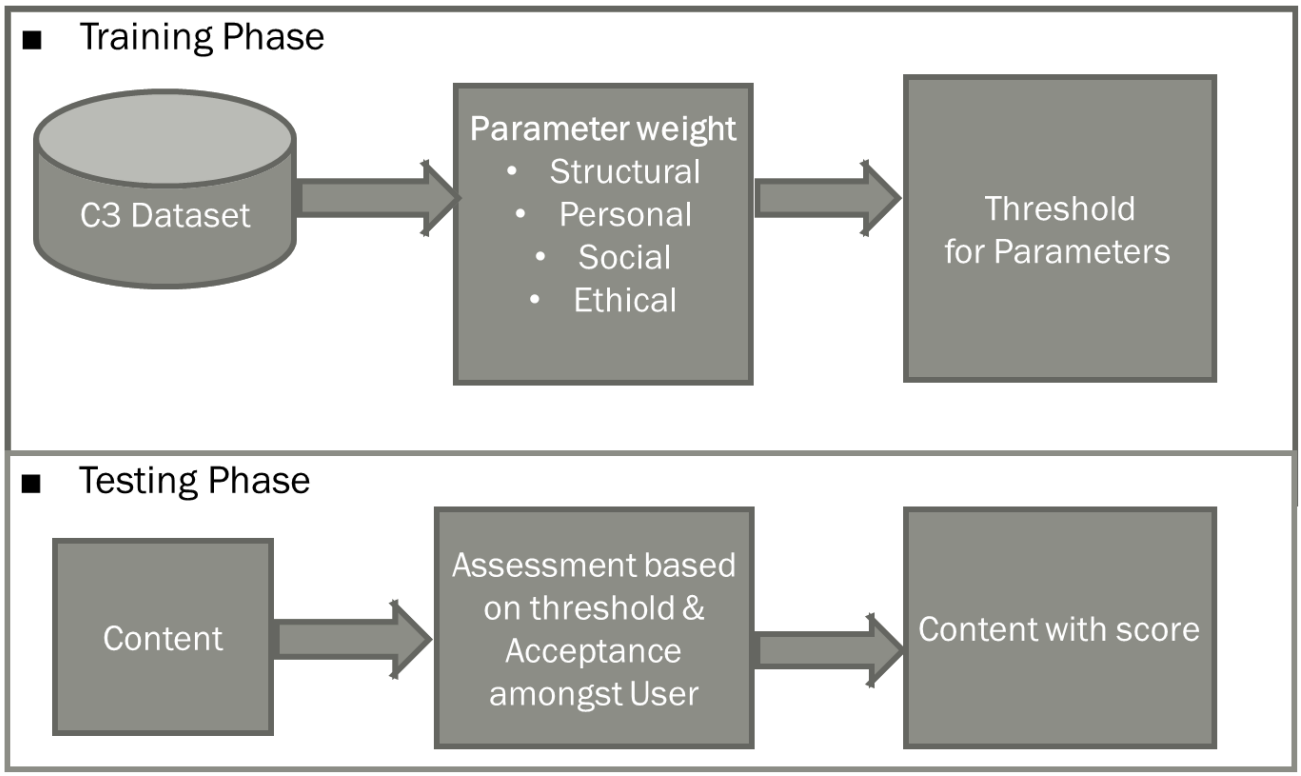

Figure 17: Architecture of the system

Training Phase:

The standard dataset is use Content Credibility Corpus for training the system

Sampled Dataset

Dataset	Web Pages	Participants	Annotations	Number of web pages used
Content Credibility Consortium (C3)	5543	2041	7071	1361

Table 2: Sampled dataset:C3

Dataset contains 5543 web pages, 2041 participants and 7071 annotations made by these users on 1361 web pages. This helps in extracting the needed features to train the system.

- Y= wx+b (Y - expected outcome, w -weight, x- input, b- bias)
- (y-b)/x=w

The training process helps in finalizing the threshold for the parameters.

Testing phase:

The threshold values act as weight assigned to each parameter, mean of these variables can indicate the performance benchmarking associated with the content. Content used for testing the performance of the system is used and the value of the result can be evaluated by using the Cronbach's Alpha Reliability Assessment tool.

Result Validation

	Mean	S.D.	1	2
p1				
p2				
p3				

Table 3: performance Measurement

Value at diagonal place indicate the maximum value, with given probability the area under the bell curve is measured with the help of Z-Table

References:

1. Detecting fake news over online social media via domain reputations and content understanding, Kuai Xu ; Feng Wang ; Haiyan Wang ; Bo Yang, Tsinghua Science and Technology (Volume: 25 , Issue: 1 , Feb. 2020)
2. Exploiting Social Review-Enhanced Convolutional Matrix Factorization for Social Recommendation, IEEE Access (Volume: 7), 2019
3. Collaborative Filtering-Based Recommendation of Online Social Voting Xiwang Yang, Chao Liang, Miao Zhao, Member, IEEE, Hongwei Wang, Hao Ding, Yong Liu, Fellow, IEEE, Yang Li, and Junlin Zhang, December 2017
2. Semantic Social Network Analysis by Cross-Domain Tensor Factorization Makoto Nakatsuji, Qingpeng Zhang, Member, IEEE, Xiaohui Lu, Bassem Makni, and James A. Hendler, Fellow, IEEE, December 2017
3. People, Technologies, and Organizations Interactions in a Social Commerce Era Nick Hajli, YichuanWang, Mina Tajvidi, and M. Sam Hajli, November 2017
4. Low-rank Multi-view Embedding Learning for Micro-video Popularity Prediction Peiguang Jing, YutingSu*, LiqiangNie, Member, IEEE, Xu Bai, Jing Liu, Member, IEEE, and Meng Wang Member, IEEE, 2017
5. Understanding content voting based on social foraging Theory, L. Xu, H. Chuan Chan, November 2017
6. Kai Shuy, Amy Slivaz, SuhangWangy, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective", 2016.

7. Sadia Afroz, Michael Brennan, and Rachel Greenstadt, “Detecting hoaxes, frauds, and deception in writing style online.” 2012.
8. Hunt Allcott and Matthew Gentzkow, “Social media and fake news in the 2016 election” Technical report, National Bureau of Economic Research, 2017.
9. MeitalBalmas, “When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism”, Communication Research, 2014.
10. Alessandro Bessi and Emilio Ferrara, “Social bots distort the 2016 us presidential election online discussion”, 2016.
11. Jonas Nygaard Blom and Kenneth Reinecke Hansen, “Click bait: Forward-reference as lure in online news headlines”, 2015.
12. Paul R Brewer, Dannagal Goldthwaite Young, and Michelle Morreale, “The impact of real news about fake news: Intertextual processes and political satire”, 2013.
13. Yimin Chen, Niall J Conroy, and Victoria L Rubin, “Misleading online content: Recognizing clickbait as false news”, 2015.
14. Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia, “Detecting automation of twitter accounts: Are you a human, bot, or cyborg” 2012.
15. Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini, “Computational fact checking from knowledge networks,” 2015.
16. M. Chang, L. Ratinov, and D. Roth. Structured Learning with Constrained Conditional Models.Machine Learning, 88(3):399{431, 2012.
17. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm.Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1{38, 1977.
18. X. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. VLDB, 2009.
19. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In WSDM, 2010.
20. K. Ganchev, J. Graca, J. Gillenwater, and B. Taskar. Posterior Regularization for Structured Latent Variable Models. Journal of Machine Learning Research, 2010.
21. P. Jorion. Risk management lessons from Long-Term Capital Management. European nancial management, 6(3):277{300, 2000.
22. Josang. Artificial reasoning with subjective logic.2nd Australian Workshop on Commonsense Reasoning, 1997.
A. Josang, S. Marsh, and S. Pope. Exploring di erent types of trust propagation. Lecture Notes in Computer Science, 3986:179, 2006.
23. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604{632, 1999.
24. H. K. Le, J. Pasternack, H. Ahmadi, M. Gupta, Y. Sun, T. Abdelzaher, J. Han, D. Roth, B. Szymanski, and S. Adali. Apollo: Towards Fact finnding in Participatory Sensing. IPSN, 2011.

B. G. Malkiel. The eficient market hypothesis and its critics. Journal of Economic Perspectives, pages 59{82, 2003.
25. Y. Nesterov and I. U. E. Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer, 2004.
26. J. Pasternack and D. Roth. Knowing What to Believe (when you already know something). In COLING, 2010.
27. J. Pasternack and D. Roth. Making Better Informed Trust Decisions with Generalized Fact-Finding. In IJCAI, 2011.
28. G. Shafer. A mathematical theory of evidence. Princeton University Press Princeton, NJ, 1976.
29. V. G. Vydiswaran, C. X. Zhai, and D. Roth. Content-driven trust propagation framework. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 974{982. ACM, 2011.
30. X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In Proc. of SIGKDD, 2007.
31. X. Yin, P. S. Yu, and J. Han. Truth Discovery with Multiple Conflicting Information Providers on the Web. IEEE Transactions on Knowledge and Data Engineering, 20(6):796{808, 2008.
32. B. Yu and M. P. Singh. Detecting deception in reputation management. Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS '03, page 73, 2003.
33. B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment, 5(6):550{561, 2012.
34. Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, 54(3), 241-251.
35. Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. Journal of the American society for information science and technology, 53(2), 134-144.
36. McKnight, D. H., & Kacmar, C. J. (2007, August). Factors and effects of information credibility. In Proceedings of the 9th International Conference on Electronic Commerce (pp. 423-432). ACM.
37. McKnight, H., & Kacmar, C. (2006, January). Factors of information credibility for an internet advice site. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS), IEEE.
38. Smith, R. E., & Vogt, C. A. (1995). The effects of integrating advertising and negative word-of-mouth communications on message processing and response. Journal of Consumer Psychology, 4(2), 133-151. [6] Wikipedia. Trustworthiness on social media. Retrieved from http://en.wikipedia.org/wiki/Social_media#Trustworthiness
39. Ball-Rokeach, S. J. (1985). The origins of individual media-system dependency a sociological framework. Communication Research, 12(4), 485-510. [8] Gaziano, C. (1988). How credible is the credibility crisis? Journalism Quarterly, 65, 267-78, 375.
40. LIVE: Verified updates. (2014). Retrieved from https://www.facebook.com/hkverified

41. Eysenbach, G. (2007). Credibility of health information and digital media: New perspectives and implications for youth.
42. Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. Communication Yearbook, 27, 293-336.
43. Newhagen, J., & Nass, C. (1989). Differential criteria for evaluating credibility of newspapers and TV news. Journalism Quarterly, 66(2), 277. [13] Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. Journal of Communication, 51(1), 52-72.
44. Armstrong, C. L., & Nelson, M. R. (2005). How newspaper sources trigger gender stereotypes. Journalism & Mass Communication Quarterly, 82(4), 820-837.
45. Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. Public Opinion Quarterly, 33(4), 563-576.
46. Burgoon, J. K., & Hale, J. L. (1984). The fundamental topoi of relational communication. Communication Monographs, 51(3), 193-214.
47. Flanagin, A. J., & Metzger, M. J. (2003). The perceived credibility of personal Web page information as influenced by the sex of the source. Computers in Human Behavior, 19(6), 683-701.
48. Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). Communication and persuasion; psychological studies of opinion change.
49. McCroskey, J. C., & Richmond, V. P. (1995). Fundamentals of human communication: An interpersonal perspective. Waveland Pr Inc.
50. Grant, A. E., Guthrie, K. K., & Ball-Rokeach, S. J. (1991). Television shopping a media system dependency perspective. Communication Research, 18(6), 773-798.
51. Critchfield, R. (1998). Credibility and web site design. [On-line]. Available: http://www.warner.edu/critchfield/hci/critchfield.html.
52. Kang, M. (2010). Measuring social media credibility: A study on a Measure of Blog Credibility. Institute for Public Relations.