

ROMCIR 2021 – April 1, 2021

# CREDIBILITY AND RELEVANCE IN INFORMATION RETRIEVAL

**GABRIELLA PASI**

UNIVERSITY OF MILANO-BICOCCA

DEPARTMENT OF INFORMATICS, SYSTEMS, AND COMMUNICATION



# THE CURRENT WEB SCENARIO

## ▶ **Web sites**

- ▶ Billions of Web pages
- ▶ E-commerce sites
- ▶ E-learning sites
- ▶ ...

## ▶ **Social Media**

- ▶ Social Networking services
- ▶ Microblogging platforms
- ▶ Media sharing networks
- ▶ ...



# INFORMATION OVERLOAD AND INFORMATION DISORDER

## ▶ **(Potential) Information overload**

- ▶ *When a decision-maker is given many sets of information, such as complexity, amount, and contradiction, the quality of its decision is decreased because of the individual's limitation of scarce resources to process all the information and optimally make the best decision [Roetzel, 2019]*

Roetzel, Peter Gordon. "Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development." *Business research* 12.2 (2019): 479-522

## ▶ **Information disorder**

- ▶ The online information ecosystem is currently polluted by various forms of information disorder, which take different forms [Wardle and Hossein, 2017]
  - ▶ Disinformation
  - ▶ Misinformation
  - ▶ Malinformation

Wardle, Claire, and Hossein Derakhshan. "Information disorder: Toward an interdisciplinary framework for research and policy making." *Council of Europe report* 27 (2017): 1-107

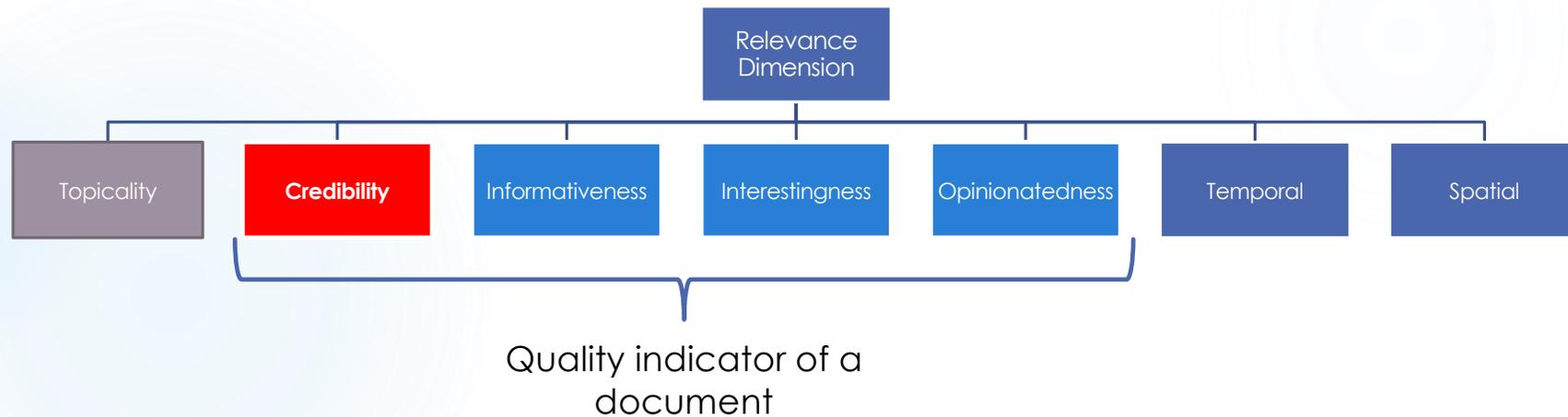
# ACCESS TO INFORMATION ITEMS

- ▶ Access to *potential information*
- ▶ Support users in **identifying information objects relevant to their needs** (to inform, i.e., reduce ignorance).
- ▶ The notion of **relevance** (*utility*) is a core notion, which is “composite” and mainly subjective.
- ▶ It has long been advocated that **relevance is a multi-dimensional property of information items**; utility to a user relies on topical relevance but it encompasses much more aspects.

# WEB SEARCH AND MULTI-DIMENSIONAL RELEVANCE

- ▶ Page popularity
  - ▶ Link-analysis
- ▶ Novelty, freshness,...
- ▶ **Learning to rank**: it accounts for different relevance factors (features)

# SOCIAL SEARCH AND MULTI-DIMENSIONAL RELEVANCE



# THE NOTION OF CREDIBILITY

## ▶ **Psychology and Sociology**

- ▶ Credibility is a perceived quality of the information receiver, based on different aspects related to:
  - ▶ The information source
  - ▶ The information diffused

## ▶ **Computer Science**

- ▶ AIM: Develop automated methods that can somehow "objectively" assess the credibility of the information or information source
  - ▶ Fake News Detection, Fake Review Detection, Rumor Detection
  - ▶ BOT detection

# INFORMATION RETRIEVAL AND CREDIBILITY

- ▶ To ensure access to information that is credible as well as relevant, it is necessary to develop **search engines that account for credibility**
- ▶ **Issues**
  - ▶ Credibility as a relevance dimension
  - ▶ Credibility as a binary or gradual property of an information item
  - ▶ At which level of granularity of a text credibility should be assessed
  - ▶ Credibility is, itself, a multi-dimensional concept:
    - ▶ Credibility of the source
    - ▶ Credibility of the message
    - ▶ Credibility of the virtual community surrounding information (in social media)

# CREDIBILITY AS A BINARY OR A GRADUAL NOTION

- ▶ In several works in the literature credibility is assessed as a **binary property of an information object**, which can be deemed either **credible** (1) or **not credible** (0)
- ▶ Other works aim to assess credibility in a **gradual** way, as a system can only estimate if an information object is credible. Lack of “objectivity”, even for human beings in certain **contexts** (e.g., reviews)
- ▶ **How to account for credibility in an IR task?** A few possible ways:
  - ▶ Assess credibility values in a numeric range and consider them as features/relevance values that concur to the overall relevance assessment
  - ▶ Return to the user an ordered list of results in which “numeric indicators of credibility” are provided for each of the returned results
  - ▶ ...

# CREDIBILITY IS A MULTI-DIMENSIONAL CONCEPT

- ▶ **Credibility of the source**

- ▶ Trustworthiness, expertise, ...

- ▶ **Credibility of the message**

- ▶ Syntactic and semantic aspects, sentiment, ...

- ▶ **Credibility of the virtual community surrounding information or information source (in social media)**

- ▶ Topological aspects of the virtual community (presence of cliques, triadic closures, ...), stance detection aspects, ...

# ANOTHER ISSUE: PERSONAL PERSPECTIVES AND TASKS

- ▶ **ISSUE:** Should credibility depend on user personal views?
  - ▶ Users may want to stay in their **filter bubble** (deliberate search for **confirmation bias**)
- ▶ **ISSUE:** Credibility and search tasks
  - ▶ **Considered Tasks** [Putri et al., 2019]
    - ▶ Disaster-related Retrieval
      - ▶ Credibility is a key concept
    - ▶ Opinion Retrieval
      - ▶ Credibility does not affect search results

# A RECURRENT ISSUE: EVALUATION

- ▶ **Lack of datasets** (and evaluation strategies and specific metrics) to assess effectiveness of:
  - ▶ Credibility prediction in social media
  - ▶ Credibility-aware Search Engines
  - ▶ Credibility-aware Recommender Systems
- ▶ Some recent datasets have been proposed in the field of information credibility assessment
  - ▶ CREDBANK - <https://github.com/compsocial/CREDBANK-data>
  - ▶ <https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php>
  - ▶ <https://iee-dataport.org/open-access/fnid-fake-news-inference-dataset>
  - ▶ ...
- ▶ What about evaluating Credibility-aware IR?
  - ▶ TREC Health Misinformation Track (2020) (<https://trec-health-misinfo.github.io/>)
  - ▶ CLEF eHealth Lab Series (2020, 2021) (<https://clefehealth.imag.fr/>)

# THE CLEF INITIATIVE eHEALTH LAB

- ▶ Retrieve and digest valid and relevant information to make health-centered decisions.
- ▶ CLEF eHealth aims to bring together researchers working on related information access topics and provide them with datasets to work with and validate the outcomes.
- ▶ **Task 1.** Multilingual Information Extraction [*\*new challenge this year*]
- ▶ **Task 2.** Consumer Health Search [*\*new challenge this year*]

# TASK 2: CONSUMER HEALTH SEARCH

## ▶ **Subtask 1: Adhoc Information Retrieval**

- ▶ evaluate IR systems ability to provide users with relevant, understandable and **credible** documents. Similarly to previous years, this subtask is centered on realistic use cases

## ▶ **Subtask 2: Weakly-supervised Information Retrieval**

- ▶ evaluate the ability of Machine Learning-based ad-hoc IR models, trained with weak supervision, to retrieve relevant documents in the health domain

## ▶ **Subtask 3: Document Credibility Assessment**

- ▶ Automatic assessment of the **credibility of information** that is disseminated online, through the Web and social media.

# TASK 2:

## AVAILABLE DATASETS

- ▶ The document collection used is the collection newly introduced in 2018, extended with **additional webpages** and **social media content**
- ▶ This collection consists of over 5 million medical webpages from selected domains acquired from the CommonCrawl and other resources
- ▶ Two **sets of topics** will be created for the task:
  - ▶ One topic set is based on discussions with multiple sclerosis and diabetes patients, the queries are manually generated by experts from established search scenarios
  - ▶ One topic set is based on use cases from discussion forums. The queries are extracted and manually selected from Google trends to best fit each use case

# OPEN ISSUES

- ▶ How to assess credibility
- ▶ To which level of granularity of an information item
- ▶ How to define datasets
- ▶ Need for new evaluation metrics

# SOME REFERENCES

- ▶ Weerkamp, Wouter, and Maarten de Rijke. "Credibility-inspired ranking for blog post retrieval." *Information Retrieval* 15.3-4 (2012): 243-277.
- ▶ Ginsca, Alexandru L., Adrian Popescu, and Mihai Lupu. "Credibility in information retrieval." *Foundations and Trends in Information Retrieval* 9.5 (2015): 355-475.
- ▶ Viviani, Marco, and Gabriella Pasi. "Credibility in social media: opinions, news, and health information—a survey." *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 7.5 (2017): e1209.
- ▶ Lioma, Christina, Jakob Grue Simonsen, and Birger Larsen. "Evaluation measures for relevance and credibility in ranked lists." *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 2017.
- ▶ Alrubaian, Majed, et al. "Credibility in online social networks: A survey." *IEEE Access* 7 (2018): 2828-2855.
- ▶ Goeuriot, Lorraine, et al. "Overview of the CLEF eHealth 2020 task 2: Consumer health search with ad hoc and spoken queries." *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

ROMCIR 2021 – April 1, 2021

Thank you for your  
attention!