December 2020

Automatic Detection of Gaps in Availability of Authoritative Online Content

Abraham Ittycheriah

Alyssa Lingad

Anne Merritt

Cong Yu

Jan Machowski

See next page for additional authors

Inventor(s)
Abraham Ittycheriah, Alyssa Lingad, Anne Merritt, Cong Yu, Jan Machowski, Kinton Cheung, Maki Sugimoto, Renato da Costa, Sarah Montgomery-Taylor, Shaan Varia, Shekhar Sharad, Srayan Datta, Tetiana Kedzierska, and You Wu

Automatic Detection of Gaps in Availability of Authoritative Online Content

ABSTRACT

This disclosure describes techniques to measure authority content gap (ACG), which represents the (lack of) authoritativeness in online content related to individual topics. The ACG metric is defined for various verticals, e.g., health, government services, legal, etc., and can be specific to region, country, language, or time period. The ACG is refreshed periodically, and it can be used in combination with other metrics relating to a content publisher. The ACG is a useful measure in various contexts, e.g., when an unpopular or obscure topic achieves sudden popularity, or when a new topic emerges. The ACG for a topic can indicate when authoritative content about such topics is unavailable and can be utilized to ameliorate the situation, e.g., by alerting content providers about the content gap.

KEYWORDS

- Content gap
- Authoritative content
- Trending topics
- Trending queries
- Search engine

BACKGROUND

A gap in the availability of reliable, authoritative online content can arise when a new topic arises and gains popularity, or a previously obscure topic for which little or no reliable content exists becomes popular. For example, emergency events such as a natural disaster or accident, other fast-moving events, trending social media posts, etc. can lead to a spike in the popularity of a topic. In the absence of availability of information about such a topic, inaccurate, unreliable, or false information can rapidly gain traction, since it may be the only information that users can find online. Such information can, in certain cases, be harmful or dangerous.

It is difficult for online entities such as content hosts/providers, social media websites, search engines, etc. to understand whether a topic has sufficient truthful, reliable information or if the information for a particular topic is predominated with unverified information. In the case of fast-moving events, contradictory online narratives about the event can emerge, e.g., via social media, various content hosts, etc. and can add to the content gap.

DESCRIPTION

This disclosure refers to gaps in content about various topics as "authority content gap" (ACG) and describes techniques to automatically measure ACG. The measured ACG can be used to identify popular topics for which little or no reliable content is available online.

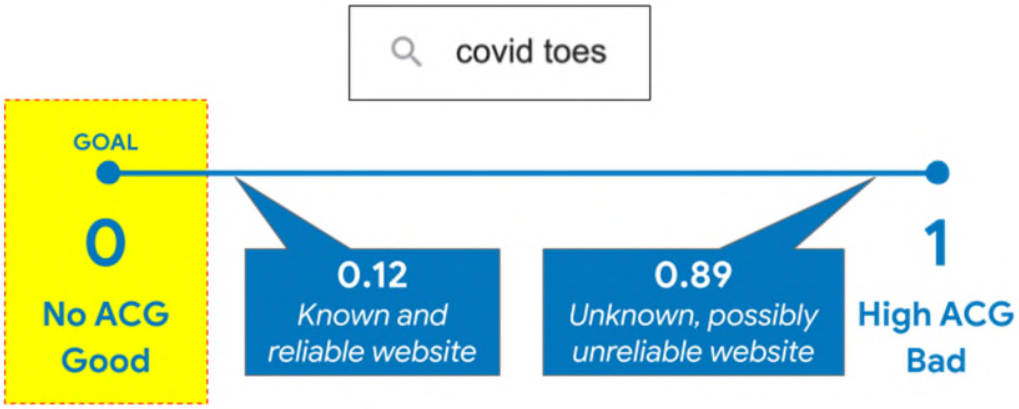

Fig. 1: Authority content gap

Authority content gap (ACG) measures the lack of authoritativeness in a search result, e.g., how authoritatively a user need is served by the results provided by a search engine in response to a search query. In the example illustrated in Fig. 1, a health-related query "covid toes" is shown. As seen in Fig.1, if the results in response to such a query are from known and reliable sources, the ACG metric is close to zero; on the other hand, if the results are mostly from unknown and possibly unreliable sources, the ACG metric is close to one.

The ACG metric can be defined for individual verticals, e.g., verticals for which authoritative voices are crucial, such as health, government services, legal, etc. The metric can be specific to region, country, language, query source (e.g., web search, video search, news search, search of digital maps, etc.,) and/or query and results interface (e.g., displayed results, spoken results, video results, etc.). ACG can be measured for different topics based on individual search results and can be refreshed frequently. ACG can be measured on a scale of zero to one, such that the ACG of search results from reliable sources is low (close to zero), while the ACG of search results from unreliable sources is high (close to one).

In some examples, to determine ACG, the following parameters are defined:

Site authority (SA): "Authoritative sites" (or sources) in a given vertical are websites with high quality-metrics that are identified by subject-matter experts as being authoritative. For example, in the health vertical in the United States, authorities can be determined to be the World Health Organization (WHO), Centers for Disease Control and Protection (CDC), etc. The authorities in the health vertical for other countries can be different, e.g., the National Health Service (NHS) in the UK, the respective ministries of health, etc. Site authority (SA) is a measure of the authority of a site, either rated by humans (as described above) or computed algorithmically.

Needs met (NM): This is a measure of whether a given document or search result meets the needs of a query. NM is either rated by humans or computed algorithmically.

User harm (UH): This is a measure of the potential harm that may be caused to the user by an inaccurate search result. UH is either rated by humans or computed algorithmically.

Authority satisfaction (AS): This is a measure of how much the user's need for information was satisfied by authoritative sources. Authority satisfaction can be computed using the formula

$$AS = SA \times NM \times UH.$$

Given the search results page for a given query q, the authority satisfaction AS_q for the query is computed automatically using the formula

$$AS_q = UH \times \sum_{i=1}^{n} W_{f_i} \times W_{p_i} \times AS_{(q,result_i)}$$

where $AS_{(q,result_i)}$ is the authority satisfaction for a particular (query, result) pair, and W_{f_i} and W_{p_i} are respectively feature weights and position weights, further defined below. Effectively, the authority satisfaction $AS_{(q,result_i)}$ for a particular (query, result) pair is computed as described above, and the overall authority satisfaction AS_q for the query q is computed as a weighted sum over $AS_{(q,result_i)}$.

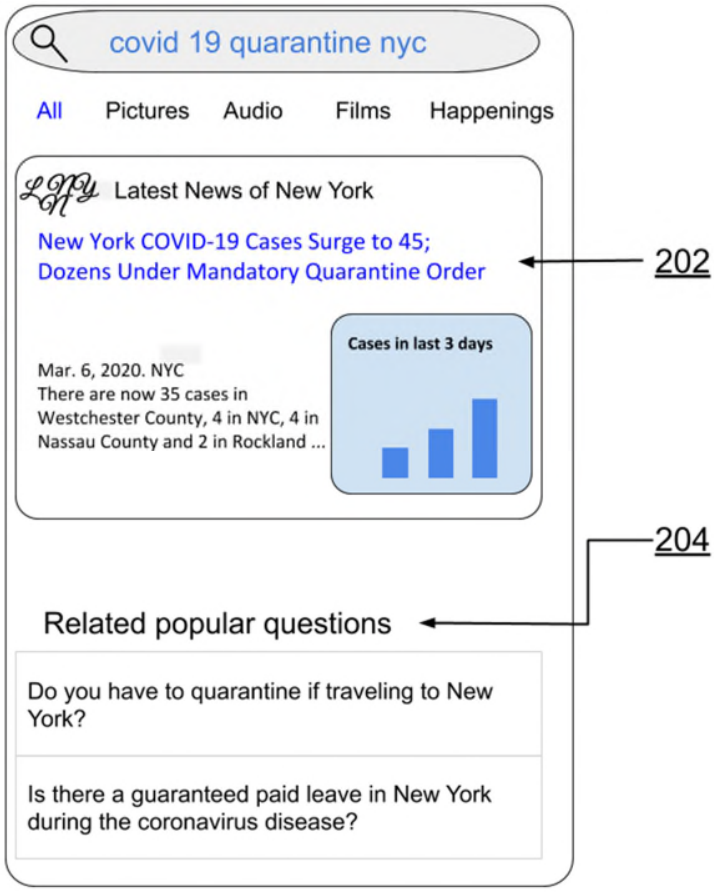

Fig. 2: Features and positions of search results

To define the feature and position weights, it is worthwhile noting again that the ACG is based on search results and does not affect search rankings. As shown in Fig. 2, a search result can be provided in many forms, e.g., as a direct result (202) that points the user to a source, as suggested commonly-asked questions and corresponding answers (204) which may be provided as a panel or as snippets, as leading current news stories, etc. The *i*th such form of search result is assigned a feature weight W_{f_i}. For example, direct results can be assigned a feature weight 1.0, while an answer to related questions can be assigned a weight 0.5.

Further, a search result has a certain position on the search results page. A position weight W_{p_i} is assigned to a result based on its position. For example, the topmost (i=1) position can be assigned a position weight of unity, and the remaining positions can be assigned corresponding descending weights. Positions beyond a certain rank can be assigned a fixed small weight, lower than the weight for all prior positions.

Once AS_q is computed for a particular query q, the authority content gap ACG_q for that query is computed using the formula

$$ACG_q = \max\left(0, \frac{AS_{zcg} - AS_q}{AS_{zcg}}\right),$$

where AS_{zcg} is the authority satisfaction for zero content-gap results, e.g., results from authoritative sources such as the CDC and WHO. The ACG_q is weighted by the logarithm of the popularity (e.g., number of searches per day) of the query to get the final authority content gap for a particular query:

$$ACG = ACG_q \times \log\left(\text{popularity}_q\right).$$

Alternatively, the ACG for a given search query can be determined by cross-referencing a search result against a list of authorities.

The described techniques can be used by any online provider to measure the gap in authoritative content related to individual topics and in individual verticals. Measurement of the ACG can enable automatic detection of topics for which authoritative content is not available. The online provider can optionally make available a list of topics or questions with high ACG to content partners (or other providers) that can generate authoritative content to fill the content gap. Further, the metric value for individual queries/topics can be monitored over time and can also be used to identify new topics that have high ACG, as queries for such topics become popular.

CONCLUSION

This disclosure describes techniques to measure authority content gap (ACG), which measures the (lack of) authoritativeness in online content related to individual topics. The ACG metric is defined for various verticals, e.g., health, government services, legal, etc., and can be specific to region, country, language, or time period. The ACG is refreshed periodically, and it can be used in combination with other metrics relating to a content publisher. The ACG is a useful measure in various contexts, e.g., when an unpopular or obscure topic achieves sudden popularity, or when a new topic emerges. The ACG for a topic can indicate when authoritative content about such topics is unavailable and can be utilized to ameliorate the situation, e.g., by alerting content providers about the content gap.