ウェブ検索結果がユーザの意見形成に及ぼす影響の調査

稲葉　健太†　　酒井　哲也†

† 早稲田大学基幹理工学部情報理工学科　〒 169-8555 東京都新宿区大久保 3-4-1
E-mail: †kinaba@ruri.waseda.jp, ††tetsuyasakai@acm.org

あらまし　情報を検索する際にウェブ検索を参考にする人は多い．ウェブ検索によってどのような情報が提供されるかは検索エンジンによって決められる．ある議論があったときにその論点をウェブ検索で集める人もいる．そんな中，たとえその検索結果が何かしらの意図である論点に偏っていてもユーザはそれが論点のすべてだと受け入れる可能性がある．本研究では、ウェブ検索結果の閲覧により，賛否両論のあるトピックに関するユーザの意見がどのように影響受けるか，特に賛否いずれかに偏った検索結果の提示をしているか否かについて簡易的な調査を行った．その調査結果を GFR と AWRF という 2 つのグループフェアな検索向け評価指標で評価した．2 つの検索エンジンの検査結果を評価したところ，2 つの検索エンジンには有意差は認められなかった．

キーワード　公平性・バイアス，検索クエリ，情報探索行動

1 はじめに

近年，情報収集手段としてインターネットを使用する人は多く，その中でも総務省によるとインターネットを使用して収集した情報を信頼する人は書籍・雑誌を上回る．[1]例えばある議論が行われているときにその論点をウェブ検索で集める人もいる．インターネットを使用して収集した情報を信頼する人がこのようにいる中で，検索結果にある論点が何らかの理由で偏っていてもユーザはそれが論点のすべてだと受け入れる可能性がある．しかしどのように情報が提供されるか，どのように情報が提示されるかは検索エンジンによるところが大きい．例えば「男女」というワードで検索をしてみるとある検索エンジンは男女という曲を 1 番上に掲載した．一方他の検索エンジンは内閣府の男女共同参画のページを 1 番上に掲載し，この男女という曲は検索結果 1 ページ目には掲載されていなかった．このように検索エンジンが選んだ結果を提示される．そんな中もしある議論の論点を検索したときに，提示される論点は検索エンジンによって決められる．この時議論の一部に限定された論点が検索結果として提示され，それをユーザが議論のすべてだと鵜呑みする可能性がある．

それ故，本研究では検索結果の提示によってユーザの意見がどのように形成されるかを調査し，偏った意見に導いているか否かを調査した．具体的には賛否両論に分かれる議論のテーマについて調べたときにどのような分布になっているかを調査した．調査結果は TREC Fair Ranking Track [1], [2] で使用されている評価指標 AWRF (Attention Weighted Rank Fairness) に加えて酒井ら [3] が提唱している指標 GFR (Group Fairness and Relevance) を使って評価した．

1：https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/html/nd125220.html

2 関連研究

2.1 情報検索における認知バイアス

Novin ら [4] は検索結果の見せ方によって認知バイアスが生じることを示した．Bailey ら [5] は検索結果画面は次第に複雑化しているという．その複雑化している検索結果画面の構成要素の一つに強調スニペットが挙げられる．Bink ら [6] は強調スニペットとバイアスとの間に相関があることを調査した．Cherumanal ら [7] は議論を検索するシステムにおいてどの立場がどの程度出現するかを測る多様な指標を分析している．Draws ら [8] は順番によるバイアスの影響があるとはいえないという．Kiesel ら [9] は音声のみで情報が提示されるシステムにおける認知バイアスについて研究している．Azzopardi [10] は認知バイアスに関する論文を 30 本以上調べ様々な領域，検索行動の様々な部分において鍵となる発見をまとめた．Gao ら [11] は公平性の条件を保持した最適化アルゴリズムの新しいフレームワークを提案している．

2.2 論点を二つに分ける

Touché at CLEF 2020 [12] で行われたタスクではオンラインでディベートをしているサイトからトピックに関連する論点を集める．その中でサイト内にある論点は PRO(賛成)/CON(反対) の 2 つの立場がある．Bondarenko ら [13] がそのタスクでの様子をまとめている．Draws ら [14] は賛成反対の 2 つの立場でしか考えてこなかった従来のものに対してユーザの意見の強さと評価論の 2 次元でユーザの考え方を示す方法を提案した．

3 ユーザ実験と評価方法

3.1 データセット

後述するユーザ実験で利用したデータセットは賛否両論あるトピックを題材に行われたタスクで使用したものを参考にした．Touché20-Argument-Retrieval-for-Controversial-Questions と

いう，Touché at CLEF 2020 のタスクのために作られたデータセット [15] の 50 のトピックから日本にいる人にもなじみがあるものを 10 個選んだ．トピックは英語の Yes/No で答えられる文で記されており，それをもとに日本語の検索クエリを作成した．検索クエリは英文の日本語訳が含まれるように作成した．作成した検索クエリを Google[2] と Bing[3] で検索し，それぞれ上位 10 位のページの和集合を今回の実験のデータセットとした．表 1 が今回のトピック，元の英文および検索クエリの対応表である．各トピックごとの和集合の平均ページ数は 18.2 と 2 つの検索エンジンにおいて重複しているページがトピックごとに 2 つ未満ということもわかった．データセット中には同じ url でも，強調スニペットを表示する text タグを含むものと含まないものの 2 種類あるものを区別した．含む url は強調スニペット中にある文章をハイライトして表示するなどその文章が際立って見える．今回そのような url の組が 3 組見つかり，人数の面ですべてにおいて同じ評価を下されていた．したがって強調スニペットによる認知の変化は観測されなかった．なお人数の面で同じ評価とは PRO と判定する人が 3 人，Neutral と判定する人が 1 人，CON と判定する人が 1 人であればその人数のみが一緒でその内訳が同一人物かどうかは定かではないということである．

3.2 ユーザ実験

ユーザ実験では日頃からネットでの検索に慣れているであろう 20 代前半の情報理工を専攻している大学生を対象とした．評価者が一番多いトピックは 7 人が評価し，少なくとも 4 人が評価している．実験協力者にエクセルシートを実験手順書と共に送付した．そのエクセルシートには一番上に今回是とした意見を表示し，その下に Google と Bing の 2 つの検索エンジンで検索した各々上位 10 個の url の和集合をランダムに表示されている．トピックも各ユーザにランダムな順番で表示されるようにした．手順書には次のようなことが主に書かれている．各 url を 30 秒ほど読み，各列最上位にある意見に対して PRO に傾くか，CON に傾くかもしくはどちらでもないかの意思表示をしてもらった．混乱を起こさないため，各ページ内にあるリンクはどれも押さないよう求めた．検索結果画面とは切り離したただの URL の羅列以外の情報をユーザは得ていない．

3.3 評価指標

3.3.1 GFR

GFR は NTCIR (NII Testbeds and Community for Information access Reserch)[4] の The NTCIR Fair Web Task [16] で使用される指標である．検索結果画面上でのユーザ分布を表す *decay* 関数，検索結果画面上の文書一致度を表す *utility* 関数と正解となる分布との比較をする分布類似性 (distribution similarity) で構成される．次のような式で表される指標である．ただし L は今回測定する順位付けされた文書のリストで，k はそのリスト中 k 番目の文書という意味である．

$$GFR_m(L) = \sum_{k=1}^{|L|} Decay^m(L,k)\,(Utility(L,k) + DistrSim^m(L,k))$$

decay 関数によって検索結果画面を上から順にみるにつれてみるユーザが減少していく様子を表し，$DistrSim^m(L,k)$ によって今回収集された分布と目指すべき分布の比較を行う．NTCIR ではクエリに対する文書の合致度の情報もあるので nDCG をもとに計算する．*decay* 関数を次のように計算した．ただし $P^{\mathrm{rel}}_{L,k} = (2^g - 1)/2^{g_{\max}}$ と記される．g は k 番目にある文書の適合レベルを示していて，$g_{\max}$ は適合レベルの最大値を意味する．

$$Decay(L,1) = P^{\mathrm{rel}}_{L,1}, \quad Decay(L,k) = p^{\mathrm{rel}}_{L,k} \prod_{j=1}^{k-1} (1 - P^{\mathrm{rel}}_{L,j})$$

今回は適合性に関する情報を集めていないので指標 GFR は直接使えない．そこで文書を読んで読者の考えに影響を与えたものを適合文書として計算することとした．トピックごとに設定した提言に PRO もしくは CON のどちらかに傾いた場合，考えに影響を及ぼしたとする．その数を適合性レベルとした．例えばあるページを読んで影響されたと考えるユーザが PRO が 3 人，Neutral が 1 人，そして CON が 1 人だったとする．この時，適合性レベルは PRO の 3 人と CON の 1 人を合わせた 4 とする．なお適合性に関する情報がない場合は RBP (Rank Biased Precision) をもとに *decay* 関数を計算することもある．これは適合性に関する情報があるときの $P^{\mathrm{rel}}_{L,k}$ が常に 0.15 であることと等しい．この Decay 関数はページの順序にのみ依存するのでページの関連度や公平さは全く寄与していない．

$$Decay(L,k) = (1-\phi)\phi^{k-1} \quad with \quad \phi = 0.85$$

Utility 関数は ERR (Expected Reciprocal Rank) [17] では $Utility(L,k) = 1/k$，iRBU (intentwise Rank-Biased Utility) [18] では $Utility(L,k) = \phi^k, (\phi = 0.99)$ で計算する．今回は ERR で計算した．

分布類似性を計算するのに欠かせない所属確率 (Group Membership Probability) は PRO が多ければ 1 に近く CON が多ければ 0 に近いような値で Neutral な意見も取り入れることを目標にした．PRO 側の所属確率は全体が 5 人で PRO 意見を持つ人が 3 人，Neutral な意見を持つ人が 1 人，そして反対意見を持つ人が 1 人いるときには $(1 * 3 + 0.5 * 1)/5 = 0.7$ とする計算式で計算した．この時 CON 側の所属確率は 0.3 である．分布類似性は今回 $\mathcal{JSD}$ (Jensen-Shannon Divergence) を使用した．名目値の対する確率間距離を測定する．今回は正解となる分布 p_* と調査対象の分布の違いをはかる．p_* は 0.5 : 0.5 の一様分布とした．これは議論の決着がついていないトピックに関して調べているので一つの論点のみを提示することは公平ではない．今回二つの論点からなるデータセットを用いたことから一様分布を正解とした．C を今回の PRO/CON の集合として p_i を i 番目の確率とする．この時 $\sum_{i \in C} p_i = 1$ である．また p^*_i を正解となる確率，今回は一様分布とする．この時 $\mathcal{JSD}$ は KLD (Kullback-Leibler Divergence) を使って次のように表される．

2：https://www.google.com/?hl=ja

3：https://www.bing.com/?cc=jp

4：https://research.nii.ac.jp/ntcir/index-ja.html

表1　トピック

トピック	英文でのトピック	検索クエリ
uniform	Should students have to wear school uniforms?	学校制服　必要性
climate	Is human activity primarily responsible for global climate change?	気候変動　人間活動　関係
game	Do violent video games contribute to youth violence?	ゲーム　暴力的　若者　影響
college	Is a college education worth it?	大学教育　価値
energy	Can alternative energy effectively replace fossil fuels?	代替エネルギー　化石燃料　転換
abortion	Should abortion be legal?	中絶　合法化
tobacco	Is vaping with e-cigarettes safe?	電子タバコ　安全性
marriage	Should gay marriage be legal?	同性婚　合法化　是非
animal	Should animals be used for scientific or commercial testing?	動物　実験利用　是非
pill	Should birth control pills be available over the counter?	アフターピル　薬局

$$\mathcal{JSD}(p, p^*) = \frac{KLD(p\|p^M) + KLD(P^*\|p^M)}{2}$$

ただし $p_i^M = (p_i + p_i^*)/2$ であり KLD は次のような式である．

$$KLD(p\|P^*) = \sum_{i \in C \text{s.t.} p_i > 0} p_i \log_2 \frac{p_i}{p_i^*}$$

式からも明らかなように足し合わせているだけなので 3 つ以上で順序がある名目を測定する際にはその距離を考慮できないが今回は PRO と CON の 2 値で測定を行ったため適しているといえる．3 つ以上で順序があるものの分布間の距離を測るためには $\mathcal{JSD}$ ではなく NMD (Normalised Match Distance) や RNOD (Root Normalised Order-aware Divergence) が挙げられる．それらの定義式は次のようであるが今回は使用していない．

$$NMD(p, p^*) = \frac{\sum_{i \in C} |cp_i - cp_i^*|}{|C| - 1}$$

ただし $cp_i = \sum_{k \leq i} p_k, \quad cp_i^* = \sum_{k \leq i} p_k^*$ である．

$$RNOD(p\|p^*) = \sqrt{\frac{OD(p\|p^*)}{|C| - 1}}$$

ただし以下の 2 式がある．

$$OD(p\|p^*) = \frac{1}{C^*} \sum_{i \in C^*} DW_i$$

$$DW_i = \sum_{j \in C} \delta_{ij} (p_j - p_j^*)^2, \quad \delta_{ij} = |i - j|$$

DW_i は正解となる確率分布からどれだけ間違った場所にあるかを測ろうとする指標である．p_j, p_j^* の差を直接測ることで実現しようとしている．また $C^* = \{i \in C | p_i^* > 0\}$ である．つまり C^* は正解となる確率分布のうち 0 を含まないもので構成されている．これら NMD と RNOD は今回のような $|C| = 2$ であるとき，等しくなることが示されている．

3.3.2　AWRF

AWRF は

$$AWRF(L) = 1 - \mathcal{JSD}(p_L^{ECE}, P_*)$$

で表現される．正解となる確率分布は GFR と同じく 0.5 : 0.5 とした．調査対象の所属確率分布も GFR と同じ計算式で求めた．ECE (Expected Cumulative Exposure) は次のような式で表される．

$$ECE(L, a_i) = \sum_{k=1}^{|L|} I(k) G(L, k, a_i) Attention(k)$$

ただし k 番目に属する文書が所属確率が正の時 $I(k) = 1$ となり，そうでないと 0 を取る．また $G(L, k, a_i)$ は文書リスト L の k 番目にある文書が値 a_i を持つ確率である．したがって $\sum_i G(L, k, a_i) = 1$ を満たす．Attention 関数は次式である．

$$Attention(k) = \frac{1}{\log_2 (k + 1)}$$

合計 10 トピックそれぞれに 20 弱の web ページが連なるリストをエクセル上で作成し，実験協力者に送付した．その際実験手順書も同封した．文書をそれぞれ評価してもらうべくトピックおよび各トピックに連なるページの順番は実験協力者によって異なる．

4　実験結果

4.1　評価結果

GFR の計算結果は次の表 2 ようになった．

表2　GFR 計算結果

	Google	Bing
uniform	0.7016	0.9169
climate	0.8705	0.9643
game	0.9924	0.9186
college	0.9383	0.9301
energy	0.9110	0.8494
abortion	0.9546	0.9405
tobacco	0.9152	0.7078
marriage	0.7215	0.9806
animal	0.9776	0.9556
pill	0.9113	0.5944
average	0.8894	0.8758

この結果をもとに対応のある t 検定を行ったところ，p 値が 0.8111 となったので Google と Bing に有意差があるとは言えないということが分かった．

AWRF の計算結果は表 3 のようになった．

表 3　AWRF の計算結果

	Google	Bing
uniform	0.891	0.927
climate	0.879	0.928
game	0.979	0.996
college	0.947	0.963
energy	0.939	0.939
abortion	0.973	0.925
tobacco	0.975	0.953
marriage	1.000	0.999
animal	0.998	0.992
pill	0.998	1.000
average	0.958	0.962

GFR と同様にこの結果を対応のある t 検定を行ったところ，p 値が 0.6348 となったので Google と Bing に有意差があるとは言えないということが分かった．また TREC Fair Ranking Track 2021 の参加者のシステムの平均 AWRF の最大値が 0.8299 となっていた．つまり今回調査した 2 つの検索エンジンの平均 AWRF はタスク参加者のシステムのそれを上回っていた．本研究と TREC ではデータもタスク設定も異なるため直接比較はできないが，あくまで平均 AWRF の絶対値としては本研究の結果が TREC と同等以上であることがわかった．2 つの検索エンジンで，GFR と AWRF の平均を比較したときに，評価指標によって優る検索エンジンが異なるということもわかった．

4.2　評価者間一致度

評価者間の一致度を示す指標である Krippendorff の α [19] (ordinal) は表 4 のようになった．評価者には提言に賛成する・反対する・わからないもしくはどちらでもないの 3 択で各文書を評価している．賛成するという意見に対してわからないもしくはどちらでもないが反対するより意見として近いと考えて計算した．評価者が評価した 2 つの検索エンジンの検索結果の和集合ではなく，2 つの検索エンジンそれぞれ 10 ページずつ合計 20 のページが各トピックで計算されている．

表 4　評価者間一致度

	評価者数	Krippendorff's α
uniform	7	0.287
climate	5	0.545
game	6	0.057
college	4	0.234
energy	4	0.119
abortion	6	-0.027
tobacco	6	0.211
marriage	6	0.397
animal	6	0.320
pill	4	0.287

Krippendorff の α は-1 から 1 の間の実数値を取り，1 に近ければ近いほど評価者間の評価が一致しているという指標で-1 に近いとその逆となる．各行に評価した値, 今回は PRO/Neutral/CON の 3 種類．各列に判定されたページを軸として，各要素があるページを PRO/Neutral/CON それぞれを判定した人数となっている行列を用いた．この時 Krippendorff の α の定義式 [20] は次式である．

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_i \sum_{j>i} o_{ij}\delta_{ij}^2}{\sum_i \sum_{j>i} e_{ij}\delta_{ij}^2}$$

ただし $n_{u\bullet} = \sum_i n_{ui}$, $n_{\bullet i} = \sum_{u|n_{u\bullet}} n_{ui}$, $n_{\bullet\bullet} = \sum_i \sum_j n_{ij}$ であり，δ_{ij} は次の式である．

$$\delta_{ij}^2 = \left(\sum_{k=i}^{j} n_{\bullet k} - \frac{n_{\bullet i} + n_{\bullet j}}{2} \right)^2$$

さらに o_{ij}, e_{ij} は次式で表される．$i \neq j$ の時

$$o_{ij} = \sum_u \frac{n_{ui} n_{uj}}{n_{u\bullet} - 1},\ e_{ij} = \frac{n_{\bullet i} n_{\bullet j}}{n_{\bullet\bullet} - 1}$$

$i = j$ の時,

$$o_{ii} = \sum_u \frac{n_{ui}(n_{ui} - 1)}{n_{u\bullet} - 1},\ e_{ii} = \frac{n_{\bullet i}(n_{\bullet i} - 1)}{n_{\bullet\bullet} - 1}$$

表 4 を見ると 0 に近い値をとっていることが多い．中絶合法化に関するトピックでは負の値をとっている．同じ文書集合を見ても人によって感じ方が違うということが再認識される．topic uniform に関しては同じ情報源をもとにした異なるページが多数上がっていたがそれでもなお評価者間で一致しなかった．多くのトピックにおいて是とした提言に関する情報がうまく得られないという声も評価者からあった．検索クエリとテーマの選び方をもう一度やる際は精査しなおす必要がある．評価者の負担を軽くするために今回各文書を少なくとも 30 秒読んでもらうことにしたが，30 秒ではあまり情報を得られず，タイトルとページ作成者が参照した情報源を見て判断したという声も評価者からあった．

5　結論と今後の課題

5.1　ま　と　め

本研究では，検索エンジンによって差別が助長されているか否かを調べるために，Google と Bing の 2 つの検索エンジンで調査した．その結果 PRO/CON の 2 つの意見に分かれたトピックにおいて，どちらか片方に大きく偏ってはいないことが分かった．また GFR と AWRF の 2 つの指標で 2 つの検索エンジンを比較した結果，GFR は Google のほうがよく，AWRF は Bing のほうが良いこともわかった．さらに対応のある t 検定の結果，2 つの検索エンジンの結果に有意差はないことも明らかとなった．

5.2　今後の課題

今回は意見生成に影響を及ぼしたものを関連性があるページとして ERR を用いて GFR を計算したが，RBP に基づく decay 関数を用いた GFR の計算も行う予定である．この減衰関数は，検索結果画面上で関連性の情報がない時に定数的に，ユーザ分

布が減衰する様子を計算する．ERR は適合レベルを用いてそれぞれの検索結果画面に合わせて計算を行う点で異なる．本実験ではエクセルシートを通してデータを収集した．今後はユーザが各ページの閲覧時間などの情報も加味した研究も可能であろう．さらに今回はこちらでクエリを設定したが，議論のあるトピックにおいてユーザがどのようなクエリを作成し論点を集めようとするかといった研究も残されている．

文　献

[1] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*, 2019.

[2] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2022.

[3] Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. 2022.

[4] Almair Novin and Eric Meyers. Making sense of conflicting science information: Exploring bias in the search engine result page. In *CHIIR '17: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 2017.

[5] Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and S.M.M. Tahaghoghi. Evaluating whole-page relevance. In *SIGIR '10: Proceedings of the 33rd international ACM SIGIR conference on Research and develop*, 2010.

[6] Markus Bink, Steven Zimmerman, and David Elsweiler. Featured snippets and their influence on users' credibility judgements. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval*, 2022.

[7] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. Evaluating fairness in argument retrieval. In *CIKM '21: Proceedings of the 30th ACM International Conference on Information Knowledge Management*, 2021.

[8] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retriev*, 2021.

[9] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. The meant, the said, and the understood: Conversational argument search and cognitive biases. In *CUI '21: Proceedings of the 3rd Conference on Conversational User Interfaces*, 2021.

[10] Leif Azzopardi. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *CHIIR '21: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 2021.

[11] Ruoyuan Gao and Chirag Shah. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *ICTIR '19: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, 2019.

[12] TOUCHE. Argument retrieval for controversial questions 2020. `https://touche.webis.de/clef20/touche20-web/argument-retrieval-for-controversial-questions`.

[13] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann abd Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2020: Argument retrieval. In *CLEF 2020 Working Notes*, 2020.

[14] Tim Draws, Oana Inel, Nava Tintarev, Christan Baden, and Benjamin Timmerman. Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval*, 2022.

[15] Potthast Martin. Touché20-argument-retrieval-for-controversial-questions. `https://zenodo.org/record/6873564#.Y5wOsXbP02w`.

[16] NTCIR. `http://sakailab.com/fairweb1/`.

[17] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of ACM CIKM 2009*, pages 621–630, 2009.

[18] Tetsuya Sakai. On the instability of diminishing return IR measures. In *Proceedings of ECIR 2021 Part I (LNCS 12656)*, pages 572–586, 2021.

[19] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (Fourth Edition)*. SAGE Publications, 2018.

[20] Tetsuya Sakai. *How to Run an Evaluation Task*, pages 71–102. Springer International Publishing, Cham, 2019.