Ranking Aggregation for Meta-Search Engine

Chunheng Jiang

November 29, 2017

Ranking aggregation is an important approach to combine information and reach an agreement between various opinions. There are many applications, such as election to select a winner from a pool of candidates based on voters' preference profile, or produce a full ranking or preference rate over a set of web pages or online movies from users' visiting log or historical ratings. The report proposed two voting rules – pairwise margin voting rule and probabilistic propagation-based voting rule. Both methods can capture the absolute and relative position information. Also, we present some evaluation results of the two methods to illustrate some good properties. We implemented a meta search engine, where the proposed pairwise margin is employed to aggregate the searching results from some individual search engines.

1 Notation

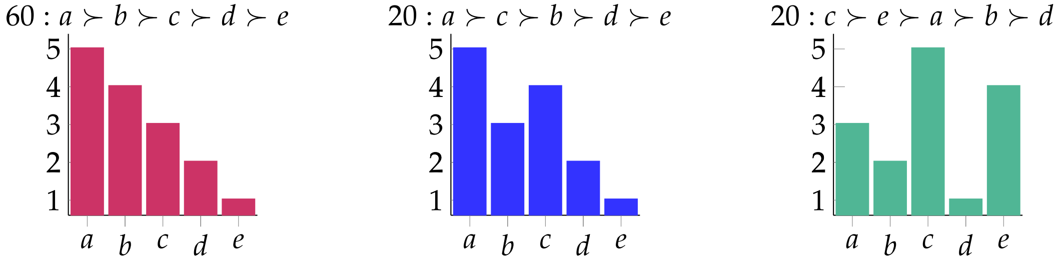
- $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ – the alternatives, candidates, agents or web pages set.
- $\Pi(\mathcal{C}) = \{\sigma_1, \sigma_2, \ldots\}$ – the set of the preference rankings (full or partial) over $\mathcal{C}$.
- $P = \{\pi_1, \ldots, \pi_n | \pi_i \in \Pi(\mathcal{C}), 1 \leq i \leq n\}$, – a preference profile from a set of voters.
- $\pi(i)$ – the position of $c_i \in \mathcal{C}$ in $\pi \in \Pi(\mathcal{C})$.
- m, n – the number of candidates and voters.

Ranking aggregation is a social welfare function r and produce an integrated result π^* from a set of known preference rankings P, i.e $\pi^* = r(P)$. The optimal Kemeny ranking is commonly used to evaluate the performance of an aggregation method, and we expect the combined result π^* is or at least very close to the optimal Kemeny ranking over P, i.e. $\pi^* \approx_{d_{KL}} \operatorname*{argmin}_{\sigma \in \Pi(\mathcal{C})} d_{KL}(\sigma, P)$.

2 Problem

We have candidates $\mathcal{C} = \{a, b, c, d, e\}$ and the preference profile of 100 voters

$$P = \{60 : a \succ b \succ c \succ d \succ e, 20 : a \succ c \succ b \succ d \succ e, 20 : c \succ e \succ a \succ b \succ d\}.$$



We denote $\pi_1 = a \succ b \succ c \succ d \succ e$, $\pi_2 = a \succ c \succ b \succ d \succ e$, and $\pi_3 = c \succ e \succ a \succ b \succ d$. Among 100 voters, 60 voters have preference ranking π_1, 20 with preference π_2 and 20 with π_3. Considering a pair of candidates a and b, their relative positions, their positional difference, and the absolute position of the preferred candidate in three preference rankings are different, as indicated in Table 1. Condorcet method could capture the relative positions of all pair of candidates. However, both the positional difference and the absolute position of the preferred candidate are ignored. Borda rule can evaluate the positional difference and the relative positions, but the absolute position of the preferred candidate is not fully considered. We expect to develop a more general voting rule to capture all three signals. The larger the positional difference between the pair of candidates, the more score the preferred candidate will be able get for winning a head-to-head competition. Furthermore, the higher the preferred candidate is ranked, the more credits it can earn.

Table 1: Absolute and Relative Positions

Positions	Condorcet	Borda
$\pi_1(a) = 1, \pi_1(b) = 2$	$a \succ b$	$s_a - s_b = 1$
$\pi_2(a) = 1, \pi_2(b) = 3$		$s_a - s_b = 2$
$\pi_3(a) = 3, \pi_3(b) = 4$		$s_a - s_b = 1$

3 Pairwise Margin Voting Rule

We proposed the pairwise margin voting rule, and tried to create a voting model that capture all three positional features for a pair of candidates: the relative position, the positional difference, and the absolute position of the preferred candidate.

Pairwise margin voting rule is a general scoring rule [1] built upon pairwise comparisons. The rule considered each pairwise competition as a zero-sum game, where a candidate earns is exactly what anothers' loss. The amount of credits that a candidate c_i gets from π for winning the pairwise competition with c_j is

$$s_\pi(i,j) = \frac{\pi(j) - \pi(i)}{\min\{\pi(i), \pi(j)\}}, \forall c_i, c_j \in \mathcal{C},$$

where $\pi(i)$ is the position of c_i in π, and $\pi(i) = 1$ is the top-most position. If candidate c_i is ranked ahead of c_j, $s_\pi(i,j) > 0$; otherwise, $s_\pi(i,j) < 0$. If ties are allowed, two tied candidates will loss and earn nothing when picked out for comparison. Besides, if c_i has no place at all, $\pi(i) = |\pi| + 1$.

Given a preference ranking π, we therefore can calculate c_i's credit by accumulating all pairwise loss and benefit, that is

$$s_i(\pi) = \sum_{1 \leq j \leq n} \frac{\pi(j) - \pi(i)}{\min\{\pi(i), \pi(j)\}}, \forall c_i \in \mathcal{C}.$$

Furthermore, candidate c_i can receive the amount $s_i = \sum_{\pi \in P(\mathcal{C})} s_i(\pi)$ of credits from a preference profile $P(\mathcal{C})$.

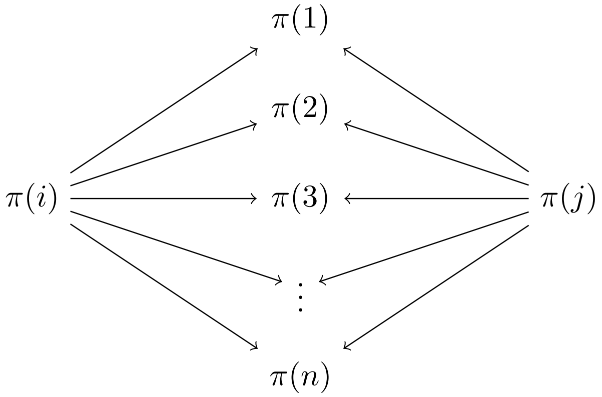


Figure 1: Monotonically increasing pairwise margin

Proposition 1. *Pairwise margin is monotonically increasing with respect to a candidate's position, i.e.*

$$\forall c_i \succ_\pi c_j, \pi \in \Pi(\mathcal{C}), s_i(\pi) > s_j(\pi).$$

With Fig. 1, here we provide a quick proof. Assuming $c_i \succ_\pi c_j$, i.e $\pi(i) < \pi(j)$, given the position $\pi(k)$ of any $c_k \in \mathcal{C}$ in π, let's see three nontrivial cases: $\pi(k) < \pi(i) < \pi(j)$, $\pi(i) < \pi(k) < \pi(j)$ and $\pi(i) < \pi(j) < \pi(k)$.

- $\pi(k) < \pi(i) < \pi(j)$: both c_i and c_j are ranked after c_k, so $\pi(k) - \pi(i) > \pi(k) - \pi(j)$,

$$\frac{\pi(k) - \pi(i)}{\min\{\pi(i), \pi(k)\}} = \frac{\pi(k) - \pi(i)}{\pi(k)} > \frac{\pi(k) - \pi(j)}{\min\{\pi(j), \pi(k)\}} = \frac{\pi(k) - \pi(j)}{\pi(k)}.$$

- $\pi(i) < \pi(k) < \pi(j)$: c_k is placed between c_i and c_j. Obviously,

$$\frac{\pi(k) - \pi(i)}{\min\{\pi(i), \pi(k)\}} = \frac{\pi(k) - \pi(i)}{\pi(i)} > 0 > \frac{\pi(k) - \pi(j)}{\min\{\pi(j), \pi(k)\}} = \frac{\pi(k) - \pi(j)}{\pi(k)}.$$

- $\pi(i) < \pi(j) < \pi(k)$: both c_i and c_j are ranked ahead of c_k, so $\frac{\pi(k)}{\pi(i)} > \frac{\pi(k)}{\pi(j)}$ and

$$\frac{\pi(k) - \pi(i)}{\min\{\pi(i), \pi(k)\}} = \frac{\pi(k) - \pi(i)}{\pi(i)} > \frac{\pi(k) - \pi(j)}{\min\{\pi(j), \pi(k)\}} = \frac{\pi(k) - \pi(j)}{\pi(j)}.$$

We end the proof.

4 Propagation-based Voting Rule

We borrow the idea of PageRank [2] developed by Segery Brin and Larry Page, the cofounder of Google, propose a propagation model to simulate a dynamic procedure in voting. The method derives from PageRank and is built upon directed weighted graphs.

4.1 PageRank

PageRank method is a probabilistic simulation of a random web surfer. Suppose the total number of web pages (or sites) is N, a web surfer is parking on page j, (s)he has to choose the next stop i to visit. There are two possible behaviors, the web surfer randomly selects another one from N pages or visits another web page by picking one hyperlinked web page contained in page j. The expected time a random web surfer visits page i is computed using the following model

$$S_i \leftarrow (1-d)/N + d \sum_{j \to i} S_j / N_j,$$

where S_i is the PageRank score of page i (or candidate c_i in voting case); d is the damping factor and also the probability of visiting another page via hyperlinks; N_j is the out-degree of vertex i. There are two terms, the first one measures the expected time of randomly jumping and the second term indicates the expected time that page i received from visiting along hyperlinks.

The score of shows a web page's importance, an item's popularity or a candidate's reputation. The importance, popularity or reputation is equally assigned to all outgoing nodes.

PageRank is built upon an unweighted directed graph, as shown in Fig 4.1 (L), where each edge is equally important. A page's score is propagated over the graph until reaching an equilibrium state.

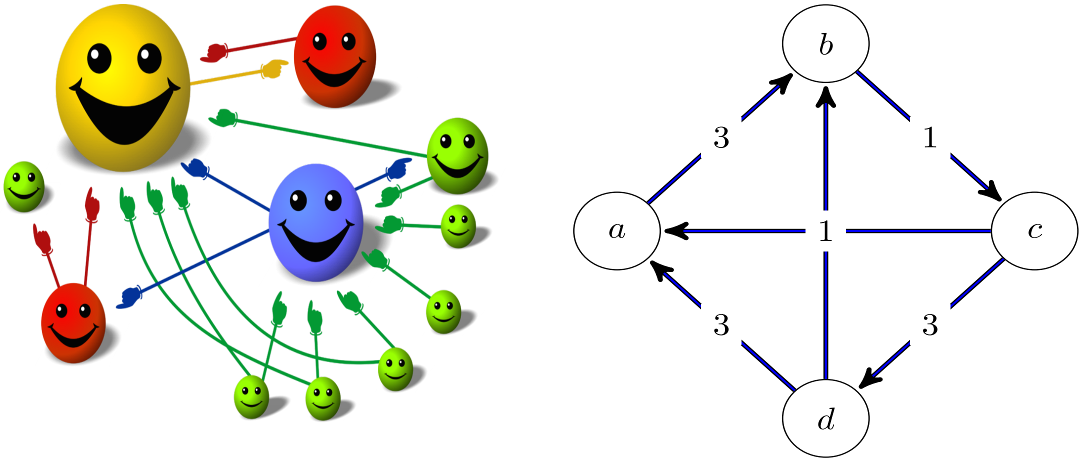


Figure 2: PageRank voting graph (L) and weighted majority graph (R)

4.2 Probabilistic Propagation

We derived from a weighted directed graph to simulate the probabilistic propagation of a web page's importance, or a candidate's popularity in election. Let $Pr(c_i|c_j)$ be the weight of the edge e_{ij} in a directed graph G or a transportation probability from c_i to c_j. We propose the following model to depict the propagation of vertices' importances

$$P_i = \sum_{j=1}^{n} Pr(c_i|c_j)P_j, \sum_{i=1}^{n} Pr(c_i|c_j) = 1, Pr(c_i|c_j) \in [0,1].$$

We can get candidates' scores by solving an eigen-system $P = AP$, where $a_{ij} = Pr(c_i|c_j)$ and $\sum\limits_{i} a_{ij} = 1$, therefore A is a stochastic matrix, which guarantees a feasible solution [3].

There are numerous ways to create a weighted directed graph and the weighted majority graph is a common way to present a preference profile - one special preference graph. This work focus on the preference graph and derives a probabilistic propagation-based voting method.

Definition 1 (Preference Graph)**.** *The preference graph of preference profile $P(C)$ is defined as a weighted directed graph whose vertices are the candidates C with edges linking all pairs of candidates. The weight w_{ij} of edge e_{ij} which links from c_j to c_i indicates the strength of $c_i \succ c_j$.*

Proposition 2. *The weighted majority graph of preference profile $P(\mathcal{C})$ is a preference graph with non-negative weights $w_{ij} = N(i \succ j) - N(j \succ i) \geq 0$ presenting that the number $N(i \succ j)$ of votes that rank c_i ahead of c_j is no less than the number $N(j \succ i)$ of votes that rank c_j ahead of c_i in the profile $P(\mathcal{C})$.*

The proposed propagation model is based on the preference graph, from which we construct a probability distribution over all pairwise preference relations. We derive from all possible pairwise elections, and formulate the following distribution

$$Pr(c_i|c_j) = \frac{\sigma(w_{ij})}{\sum\limits_{k=1}^{n} \sigma(w_{kj})}, \forall c_i, c_j \in \mathcal{C},$$

where $\sigma(x) = 1/[1 + e^{-x}]$. Let $w_{ij} = s_i - s_j$, we have a propagation-based voting rule based on pairwise margin voting rule.

5 Evaluation and Application

We enumerated all possible preference profiles of $m = 3$ candidates from $n = \{5, 7, 9, \ldots\}$ voters, and conducted two kinds of evaluations: (a) the similarities of the proposed methods to Kemeny-Young method, (b) the satisfiabilities to some popular fairness criteria using the approach suggested by Lirong Xia [4]. Moreover, we also implemented a meta search engine with Google, Yahoo!, Ask, Baidu, Bing and Blekko as its engine members, and employed the proposed pairwise margin voting rule to aggregate the top ranked search results from all these individual search engines.

As indicated in Fig. 3(L), Condorcet method is much similar to the Kemeny-Young method than all the other voting rules. Based on the observation, we have to admit that the proposed method is not suitable for ranking aggregation based on its similarity to Kemeny-Young rule. Pairwise margin and Borda rule perform similarly in searching the nearest ranking to a given preference profile, because they behave similarly in selection a winner (see Fig. 3 (R)).

An independent general search engine contains four primary components [2, 3]: web crawler (collect web pages), documents indexer(create forward document-term indexing and inverted index of terms), ranker (analysis documents' signals and rank them using various ranking algorithms) and searching terminal (parse queries and communicate with other three components for response). Researches shown that the most well known search engines have low overlap [5] in searching results and their ranking performances are different as well. It's valuable to combine results from multiple sources can provide more diverse coverage, and to some extend could reduce the distraction from advertises. There are also disadvantages in a meta search engine. The most obvious one is that most

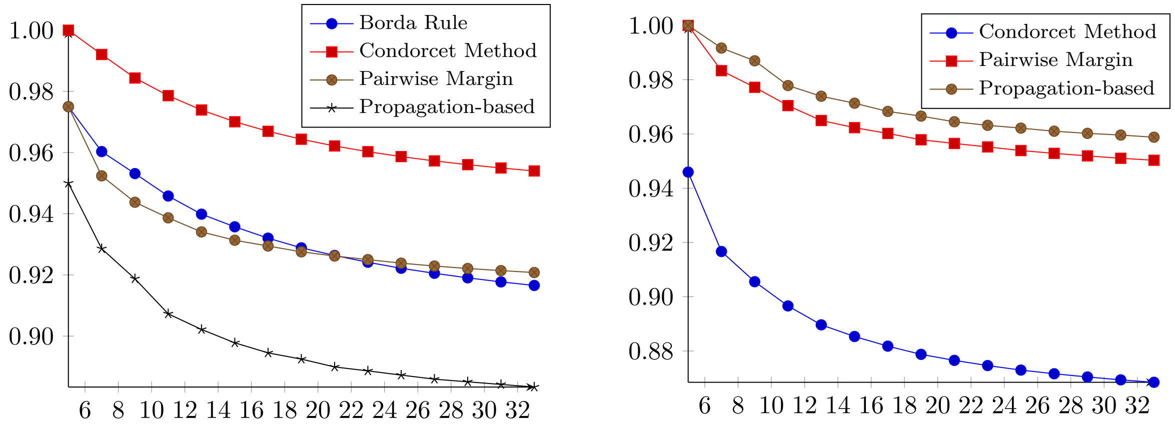


Figure 3: Comparisons of pairwise margin, propagation-based voting rule, Borda rule and Condorcet method in terms of similarity to Kemeny-Young (L); Comparisons of pairwise margin, propagation-based voting rule and Condorcet method in terms of similarity to Borda rule (R).

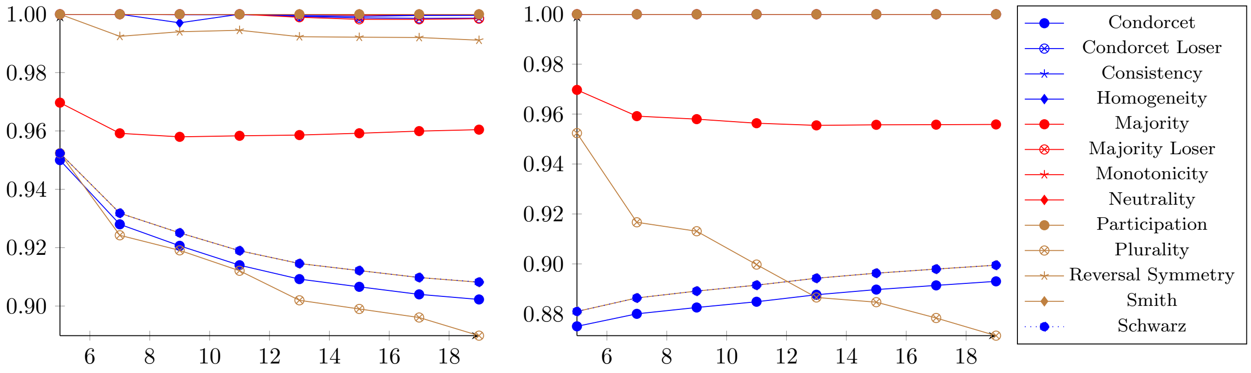
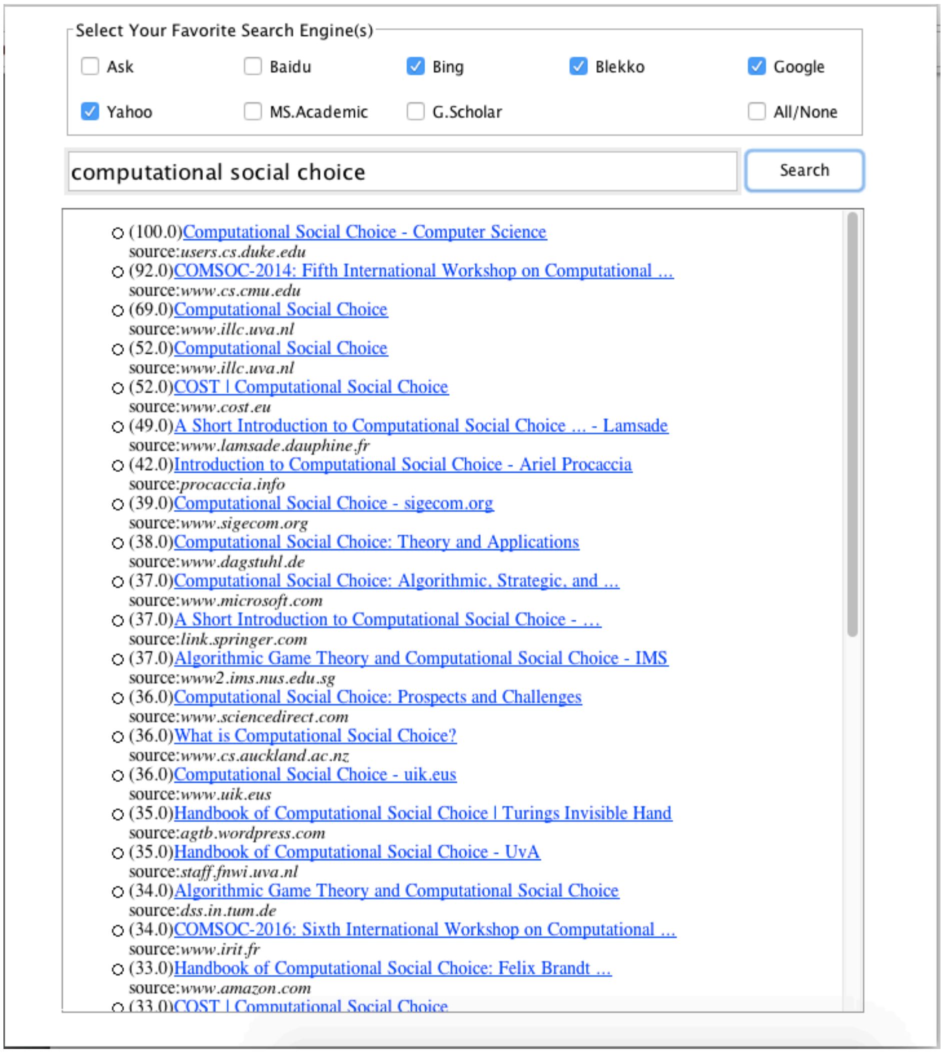


Figure 4: Comparison between propagation-based voting rule (L) and Borda rule (R) in terms of satisfiability to popular fairness criteria

meta search engine do not have independent web database, and therefore require longer waiting time.

The implemented engine is not optimized and only provide basic searching functionality (see Fig. 5). There are lots of work to do to provide real-time service, including deduplication detection, searching cache, query spelling correction, keywords highlight, etc.



Select Your Favorite Search Engine(s)

☐ Ask ☐ Baidu ☑ Bing ☑ Blekko ☑ Google

☑ Yahoo ☐ MS.Academic ☐ G.Scholar ☐ All/None

computational social choice

- ○ (100.0)Computational Social Choice - Computer Science
  source:*users.cs.duke.edu*
- ○ (92.0)COMSOC-2014: Fifth International Workshop on Computational ...
  source:*www.cs.cmu.edu*
- ○ (69.0)Computational Social Choice
  source:*www.illc.uva.nl*
- ○ (52.0)Computational Social Choice
  source:*www.illc.uva.nl*
- ○ (52.0)COST | Computational Social Choice
  source:*www.cost.eu*
- ○ (49.0)A Short Introduction to Computational Social Choice ... - Lamsade
  source:*www.lamsade.dauphine.fr*
- ○ (42.0)Introduction to Computational Social Choice - Ariel Procaccia
  source:*procaccia.info*
- ○ (39.0)Computational Social Choice - sigecom.org
  source:*www.sigecom.org*
- ○ (38.0)Computational Social Choice: Theory and Applications
  source:*www.dagstuhl.de*
- ○ (37.0)Computational Social Choice: Algorithmic, Strategic, and ...
  source:*www.microsoft.com*
- ○ (37.0)A Short Introduction to Computational Social Choice - ...
  source:*link.springer.com*
- ○ (37.0)Algorithmic Game Theory and Computational Social Choice - IMS
  source:*www2.ims.nus.edu.sg*
- ○ (36.0)Computational Social Choice: Prospects and Challenges
  source:*www.sciencedirect.com*
- ○ (36.0)What is Computational Social Choice?
  source:*www.cs.auckland.ac.nz*
- ○ (36.0)Computational Social Choice - uik.eus
  source:*www.uik.eus*
- ○ (35.0)Handbook of Computational Social Choice | Turings Invisible Hand
  source:*agtb.wordpress.com*
- ○ (35.0)Handbook of Computational Social Choice - UvA
  source:*staff.fnwi.uva.nl*
- ○ (34.0)Algorithmic Game Theory and Computational Social Choice
  source:*dss.in.tum.de*
- ○ (34.0)COMSOC-2016: Sixth International Workshop on Computational ...
  source:*www.irit.fr*
- ○ (33.0)Handbook of Computational Social Choice: Felix Brandt ...
  source:*www.amazon.com*
- ○ (33.0)COST | Computational Social Choice

Figure 5: A screenshot of the meta search engine implemented in Java

6 Conclusion and Future Work

We proposed pairwise margin voting rule to capture all three important positional features: the relative position, the positional difference, and the absolute position of the preferred candidate in head-to-head competition. It has higher satisfiability than Borda rule to the Condorcet criterion, a very good fairness criterion. Furthermore, we also create a general probabilistic propagation-based voting rule based on the idea of PageRank. It can efficiently give the scores of candidates iteratively and presents a flexible approach to formulate pairwise preference relation.

The proposed methods are not fully analyzed and require further investigation and comparisons with other positional scoring rules and pairwise Condorcet method.

References

[1] Lirong Xia and Vincent Conitzer. Generalized scoring rules and the frequency of coalitional manipulability. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 109–118. ACM, 2008.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[3] Amy N Langville and Carl D Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[4] Lirong Xia. Designing social choice mechanisms using machine learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 471–474. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[5] Amanda Spink, Bernard J Jansen, Vinish Kathuria, and Sherry Koshman. Overlap among major web search engines. *Internet Research*, 16(4):419–426, 2006.