Fairness in Rankings

Meike Zehlike[1]

Abstract: Künstliche Intelligenz und selbst-lernende Systeme spielen eine immer größer werdende Rolle in unserem Alltag. Rankings stellen dabei das wesentliche Instrument unserer Onlinesuche nach Inhalten, Produkten, Freizeitaktivitäten und relevanten Personen dar. Die Reihenfolge der Suchergebnisse bestimmt somit nicht nur die Zufriedenheit der Suchenden, sondern auch die Chancen der Sortierten auf Bildung, ökonomischen und sogar sozialen Erfolg. Diskriminierende Rankings erzeugen nicht nur unmittelbare Nachteile, sondern führen auch zu nachteil-verstärkenden Feedbackschleifen. Die vorliegende Arbeit adressiert drei wichtige Herausforderungen, die im Kontext algorithmischer Diskriminierung durch Ranking-Systeme auftreten: Die ethischen Ziele verschiedener Ranking-Situationen müssen mit denjenigen übereinstimmen, die in Ranking-Algorithmen implizit kodiert sind. Zweitens müssen ethische Wertesysteme in Mathematik und Algorithmen zu übersetzt werden, um Ranking-Algorithmen zur Vermeidung von Diskriminierung bereitzustellen. Drittens sollten diese Methoden einem breiten Publikum zugänglich sein, das sowohl Programmiererinnen, als auch Juristinnen und Politikerinnen umfasst.

1 Einführung

Künstliche Intelligenz und selbst-lernende Systeme, die ihr Verhalten aufgrund vergangener Entscheidungen und historischer Daten adaptieren, spielen eine immer größer werdende Rolle in unserem Alltag.Wir sind umgeben von einer großen Zahl algorithmischer Entscheidungshilfen, sowie einer stetig wachsenden Zahl algorithmischer Entscheidungssysteme. Rankings und sortierte Listen von Suchergebnissen stellen dabei das wesentliche Instrument unserer Onlinesuche nach Inhalten, Produkten, Freizeitaktivitäten und relevanten Personen dar. Aus diesem Grund bestimmt die Reihenfolge der Suchergebnisse nicht nur die Zufriedenheit der Suchenden, sondern auch die Chancen der Sortierten auf Bildung, ökonomischen und sozialen Erfolg. Diskriminierung und systematische Abwertung bestimmter Personengruppen in Rankings erzeugen nicht nur unmittelbare Nachteile für die Sortierten, sondern zusätzlich negative Feedbackschleifen, die die Lage der ohnehin schon Benachteiligten noch weiter verschlechtert. Betrachten wir als Beispiel eine Online-Plattform bei der sowohl Arbeitsuchende, als auch Recruiterinnen[2] sich anmelden um das passende Gegenstück zu ihrer Suche zu finden. Nehmen wir an eine Recruiterin sucht nach einer geeigneten Kandidatin für eine Position im Bereich Software Engineering. Durch vielseitige Arten von Bias hat der Ranking-Algorithmus gelernt, Frauen für diesen Bereich als weniger relevant

[1] Humboldt Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, Institut für Informatik, Rudower Chaussee 25, 12489 Berlin, Deutschland zehlike@gmail.com

[2] Ich verwende in dieser Zusammenfassung ausschließlich das generische Femininum, beziehe mich aber natürlich nicht nur auf Frauen, sondern auf alle Personen.

CC BY SA

einzuschätzen als Männer. Aus diesem Grund wird Klara Musterfrau trotz einschlägigen Studiums mit guten Noten an das Ende der Liste verbannt, wo sie für die Recruiterin praktisch unsichtbar wird. Sie wird nicht kontaktiert und erhält weniger Angebote, was zu weniger Wahlmöglichkeiten mit weniger Entwicklungspotential führt, was zu weniger relevanter Berufserfahrung führt, was zu geringerer Vermittelbarkeit führt. Rankings bilden also nicht einfach nur eine (ohnehin diskriminierende) Wirklichkeit ab, sie erschaffen eine neue Realität, die vorhandene Diskriminierung noch verschärft [ON17]. Wissenschaft und Politik sorgen sich aus diesem Grund mehr und mehr um systematische Diskriminierung und Bias durch selbst-lernende Systeme.

Um der Diskriminierung im Kontext von Rankings Herr zu werden, sind folgende drei Herausforderungen zu addressieren: Zunächst müssen wir die ethischen Eigenschaften und moralischen Ziele verschiedener Situationen erarbeiten, in denen Rankings eingesetzt werden [ZYS22a; ZYS22b]. Diese Ziele sollten mit den ethischen Werten der Algorithmen übereinstimmen, die zur Vermeidung von diskriminierenden Rankings Anwendung finden. Zweitens ist es notwendig, ethische Wertesysteme in Mathematik und Algorithmen zu übersetzen, um sämtliche moralischen Ziele bedienen zu können [ZC20; Ze17; Ze22b]. Drittens sollten diese Methoden einem breiten Publikum zugänglich sein, das sowohl Programmiererinnen, als auch Juristinnen und Politikerinnen umfasst. Die vorliegende Arbeit [Ze22a] widmet sich der Lösung dieser Fragestellungen auf folgende Art:

Fünf Konzepte zur Klassifizierung von Fairness: Wir untersuchen alle wichtigen Definitionen von Fairness für Rankings hinsichtlich ihrer ethischen Werte und moralischen Ziele, sowohl jene aus der Literatur, als auch die in dieser Arbeit neu Eingeführten. Dafür stellen wir fünf Dimensionen von Fairness vor und eine systematische Klassifizierung aller in dieser Arbeit präsentierten Methoden auf.

Faire Ranking Algorithmen: Wir präsentieren zwei neue Ranking-Methoden gegen indirekte Diskriminierung, wovon eine Methode die Zielfunktion eines Learning-to-Rank Algorithmus um eine Fairness-Komponente erweitert und die andere ein bereits vorhandenes Ranking nach Fairnesskriterien neu sortiert.

Eine Open-Source API: Wir implementieren unsere Fair-Ranking Methoden in die erste open-source Bibliothek für faire Suchergebnisse, einerseits als Python- und Java-Pakete, andererseits als Plugin für die bekannte Such-API "Elasticsearch."

2 Grundlagen und Problemstellung

Rankings sind ein zentraler Bestandteil vieler Anwendungen im Bereich Information Retrieval, seien es Suchmaschinen oder Empfehlungssysteme. Sie werden erzeugt, indem alle Dokumente[3] aus einer Ergebnismenge hinsichtlich ihrer geschätzten Relevanz für

[3] Dokumente dient hier als generische Bezeichnung für die Elemente der Ergebnismenge. Ein Dokument kann also sowohl ein echtes Dokument sein, als auch ein Produkt oder das Profil einer Kandidatin.

die Nutzerin absteigend sortiert werden. Somit kann die Nutzerin sich direkt denjenigen Dokumenten widmen, die wahrscheinlich am interessanten für sie sind, anstatt sich durch die potentiell riesige Menge aller als "relevant" markierten Dokumente zu arbeiten.

Die Qualität $U(\tau)$ eines Rankers f wird häufig mittels Normalized Discounted Cumulated Gain (NDCG) [JK02] gemessen. Dabei wird die Relevanz y eines Dokuments $d_i = \tau(j)$ in Relation zu seiner Position j im von f erzeugten Ranking τ gesetzt: $U(\tau) = \text{NDCG} = \frac{1}{\text{IDCG}} \sum_{j=1}^{k} \frac{y_{\tau(j)}}{\log_2(j+1)}$. Diese Gleichung trägt der Tatsache Rechnung, dass die Sichtbarkeit innerhalb der ersten k Positionen im Ranking rapide abnimmt, was als *Positionsbias* [Jo17] bekannt ist. Studien zeigen, dass die Aufmerksamkeit, die Nutzerinnen den Dokumenten in ihren Suchergebnissen schenken, exponentiell mit jeder Position sinkt.

Geschützte Merkmale und demografische Gruppen. Um Diskriminierung in Rankings erkennen und korrigieren zu können, benötigen wir zunächst das Konzept einer *demografischen Gruppe*, deren Zugehörigkeit über die Anwesenheit bestimmter *geschützter Merkmale*, wie z.B. Gender, Behindertenstatus oder Migrationshintergrund, definiert wird. Dabei können Kandidatinnen[4] gleichzeitig mehreren Gruppen angehören (z.B. Frauen mit Migrationshintergrund). Solche Menschen erfahren einerseits stärkere Benachteiligung als Personen, die nur einer Gruppe angehören. Andererseits folgt die Diskriminierung meist völlig anderen Mustern als solche, die die einzelnen Gruppen betrifft, d.h. man kann nicht einfach davon ausgehen, dass das Ausmaß der Diskriminierung linear mit der Anzahl an Gruppenzugehörigkeiten zunimmt. Dieses Phänomen wird als *Intersektionalität* [Cr90] bezeichnet.

Fairness und Diskriminierung durch ungleiche Sichtbarkeit im Ranking. Wir verstehen Diskriminierung bzw. Unfairness in Rankings als eine *systematische Ungleichheit in der Sichtbarkeit*, sowohl für die Einzelne, als auch für Gruppen von Kandidatinnen. Diese Interpretation liegt in der direkten Abhängigkeit von Sichtbarkeit mit wirtschaftlichem, beruflichem und sozialem Erfolg derjenigen, die gerankt werden, begründet. Wenn eine Person oder auch eine bestimmte demografische Gruppe systematisch auf die hintersten Plätze verbannt wird, sinkt ihre Chance, z.B. auf ein Vorstellungsgespräch, wegen des Positionsbias' und der exponentiell sinkenden Sichtbarkeit gen Null. Dabei werden ungleiche und unfaire Verteilungen von Sichtbarkeit zunächst durch den Positionsbias hervorgerufen, insbesondere dann, wenn Kandidatinnen sich qualitativ sehr ähnlich sind, jedoch eine (ähnlich qualifizierte) Kandidatin im Ranking naturgemäß vor der anderen kommen muss. Hinzu kommen Bias und Diskriminierung aus der Gesellschaft, die ihren Weg in Daten und Algorithmen finden.

Die vorliegende Arbeit präsentiert zwei Methoden zur Reduzierung von Diskriminierung in Rankings, die sich als richtungsweisend für das gesamte Feld erwiesen haben: FA*IR [Ze17; Ze22b; ZSC20] gehört zu den ersten fairen Ranking-Algorithmen überhaupt und begründete

[4] Ab diesem Kapitel verwende ich die Begriffe 'Dokument' und 'Kandidatin' synonym für die Elemente eines Rankings, um zu verdeutlichen, dass sich hinter dem abstrakten Konzept des Dokuments ein realer Mensch und sein Schicksal verbirgt.

den Bereich der probabilistischen Definitionen fairer Rankings. Mit DELTR [ZC20] stelle ich den ersten fairen Lernalgorithmus für Rankings vor. Seine Fairness-Definition gehört zum Paradigma der expositionsbasierten Definitionen. Letzteres Paradigma misst Sichtbarkeit als die zu erwartende Aufmerksamkeit, die eine Nutzerin einem Dokument beimessen wird, was eng mit der Definition des Positionsbias zusammenhängt. Das probabilistische Paradigma stellt die Frage, ob ein Ranking mit einer hinreichenden Wahrscheinlichkeit durch einen fairen statistischen Prozess (z.B. durch würfeln) erzeugt wurde. Im Gegensatz zum expositionsbasierten Paradigma wird hier jede Position einzeln überprüft, sodass es nicht möglich ist eventuell vorhandene lokale Diskriminierung in den ersten Positionen in späteren Positionen zu korrigieren.

3 Fünf Dimensionen für Fairness in Ranking-Algorithmen

Um die verschiedenen formellen Definitionen von Fairness für Rankings richtig einordnen und in Beziehung setzen zu können, brauchen wir ein gemeinsames Verständnis von Bias und Fairness sowie deren moralischer Nuancen einerseits, und von den verschiedenen Perspektiven hinsichtlich derer algorithmische Fairness betrachtet werden kann andererseits. Im Vergleich stellt man fest, dass faire Algorithmen sich stark darin unterscheiden, wie sie Kandidatinnen repräsentieren (z.B., ob sie nur ein geschütztes Merkmal behandeln oder mehrere), welche Art Bias sie zu erkennen und verhindern suchen, wie sie mit Trade-offs zwischen Fairness und Relevanz umgehen, und an welcher Stelle innerhalb der Machine-Learning-Pipeline sie ansetzen. Diese technischen und operationellen Entscheidungen stellen normative Entscheidungen dar, die die vorliegende Arbeit durch sorgfältige Analyse der Beziehung von ethischem Konzept zu mathematischer Definition sichtbar macht.

Modellierung der Gruppenstrukur. Beinahe alle Methoden, die im Bereich faire Rankings bisher vorgeschlagen wurden, nehmen geschützte Merkmale als kategorisch (also nicht kontinuierlich) an. Viele Methoden können lediglich ein einziges Merkmal gleichzeitig behandeln, und viele dieser Methoden haben die zusätzliche Einschränkung, dass dieses Attribut binär sein muss (d.h., es gibt genau eine geschützte und eine nicht-geschützte Gruppe). Mit der Zeit wurden einige dieser Arbeiten auf mehrere Merkmale generalisiert, unter anderem eine der beiden Methoden dieser Dissertation. Nur solche Ansätze mit der Möglichkeit mehr als eine geschützte Gruppe gleichzeitig in Betracht zu ziehen sind in der Lage intersektioneller Diskriminierung zu begegnen. Somit birgt die mathematische Beschränkung auf nur eine mögliche geschützte eine normative Komponente, da sie Intersektionalität ignoriert.

Typen von Bias. Der Begriff *Computer Bias* stammt von Friedman und Nissenbaum [FN96], die ein System als ‘gebiast’ deklarieren, wenn es bestimmte Individuen oder Gruppen *systematisch* ungleich und gleichzeitig *unfair* behandelt. Dabei meint unfaire Behandlung entweder das *Versagen einer Leistung* oder das *Zuweisen eines unerwünschten Resultats* auf Basis von *unangebrachten* oder *unvernünftigen* Gründen. Bias hat also immer eine systematische und eine unfaire Komponente. Weiterhin definieren Friedman und Nissenbaum

drei allgemeine Typen von Bias: (1) *Bestehender Bias:* Dieser Typ enthält jeglichen Bias der unabhängig vom System existiert und dessen Ursprung in der Gesellschaft und ihren Strukturen zu finden ist. (2) *Technischer Bias:* Dieser Typ bezieht sich auf sämtliche Biases, die durch technische Grenzen und Beschränkungen hervorgerufen werden. Der Positionsbias fällt in diese Kategorie. (3) *Entstehender Bias:* Dieser Typ beschreibt Biases, die erst während der aktiven Nutzung eines Systems entstehen, entweder weil die antizipierten Nutzerinnen nicht den realen entsprechen, oder weil sich gesellschaftliche Konzepte weiterentwickelt haben.

Ansatzpunkt in der Machine-Learning-Pipeline. Wie schon der vorherige Abschnitt erahnen lässt, tritt Bias in allen Phasen der Machine-Learning-Pipeline auf, akkumuliert sich und gewinnt häufig auf seinem Weg an Bedeutsamkeit. Es wurden für jede Stufe Methoden vorgeschlagen, die, je nachdem ob sie vor, während oder nach dem Modell-Training ansetzen, als *Pre-, In-* oder *Post-processing* bezeichnet werden.

Weltanschauung. Diese Dimension zur Klassifizierung von fairen Rankern behandelt die moralischen Ziele bei der Bias-Minderung und verdeutlicht, welche Eigenschaften ein faires Ranking bedingen. Friedler et al. [FSV16] definieren drei für das maschinelle Lernen relevante metrische Räume: den Konzeptraum (*CS*), den beobachtbaren Raum (*OS*) und den Entscheidungsraum (*DS*). Die Beziehung zwischen diesen Räumen hängt nun maßgeblich vom Wertesystem der jeweiligen Entwicklerin und Nutzerin ab und behandeln die Frage, ob sich die tatsächlichen Stärken und Schwächen eines Menschen mittels beobachtbarer und maschinenlesbarer Daten abbilden lassen. Der Weltanschauung 'What you see is what you get' (WYSIWYG) liegt die Annahme zu Grunde, dass eine Abbildung von *CS* nach *OS* nur sehr geringe Verzerrungen in den Daten erzeugt. Die Anschauung 'We are all equal' (WAE) hingegen geht davon aus, dass alle beobachteten Unterschiede zwischen den Verteilungen verschiedener Gruppen auf fehlerhafte Messungen und historische Diskriminierung zurückgehen.

Verständnis von Chancengleichheit. Chancengleichheit ist ein Konzept aus der politischen Philosophie, das sich mit der Frage nach der gerechten Verteilung von gesellschaftlichen Gütern und Positionen beschäftigt. Heidari et al. [He19] stellen vier verschiedene Subtypen vor, von denen ich hier auf zwei näher eingehe: (1) *Formelle Chancengleichheit:* in diesem Paradigma werden gesellschaftlich-wünschenswerte Güter und Positionen ausschließlich aufgrund von Merkmalen vergeben, die für die Ausübung der Position relevant sind (z.B. darf niemandem aufgrund ihrer Hautfarbe der Zugang verwehrt werden). Diejenige, die am besten für die Position geeignet ist, erhält sie. (2) *Substanzielle Chancengleichheit (speziell Luck-Egalitarianism):* in diesem Paradigma werden Güter und Positionen ebenfalls nach relevanter Qualifikation vergeben, jedoch wird hier zusätzlich gefordert, dass die Möglichkeit die notwendige Qualifikation zu erwerben mit in Betracht gezogen wird. Dies erkennt an, dass unterschiedliche Startbedingungen einen großen Einfluss auf die Qualifizierungsmöglichkeiten eines Menschen haben. Luck-Egalitarianism im Speziellen fordert hierbei, nur Individuen einer Gruppe miteinander zu vergleichen.

In den nächsten Abschnitten werde ich die vorgestellten Methoden anhand dieser fünf Dimensionen vergleichen und sie dadurch miteinander in Beziehung setzen.

4 FA*IR: Ein fairer top-k Ranker

Der Algorithmus FA*IR beruht auf einem statistischen Signifikanztest und kann einerseits ein gegebenes Ranking auf Fairness überprüfen, andererseits ein qualitativ hochwertiges und gleichzeitig faires Ranking mit k Kandidatinnen erzeugen. In einem solchen Ranking, nennen wir es τ, darf der Anteil an Kandidatinnen einer geschützten Gruppe an *keiner Position* unter ein gegebenes Minimum fallen. Dieses Kriterium nennen wir *Ranked Group Fairness*. Mathematisch ausgedrückt vergleicht dieses Kriterium die Anzahl geschützter Kandidatinnen in jedem Ranking-Präfix mit deren erwarteter Anzahl, wären sie durch ein rein stochastischen Prozess (in unserem Fall die Multinomialverteilung, d.h. an jeder Position wird gewürfelt) ausgewählt worden. Unter einem Ranking-Präfix $\tau_{1..j}$ sind jeweils die ersten j Positionen in τ zu verstehen. Der Algorithmus erhält als Input die Rankinglänge k, sowie den Minimalanteil p_i an Kandidatinnen für jede geschütze Gruppe $G_i \in \mathbf{G}$ (d.h. wir würfeln mit Wahrscheinlichkeit p_i eine Kandidatin aus Gruppe G_i). Im Anschluss ermittelt er für jede Gruppe die Anzahl τ_i an geschützten Kandidatinnen aus Gruppe G_i im Ranking τ. Den Vektor, der diese Anzahlen enthält, nennen wir $\tau_{\mathbf{G}}$.

Definition 1 (Fair Representation Condition) *Sei n die Anzahl an Versuchen, von denen jeder mit Wahrscheinlichkeit p_i zu Ereignis E_i führt. Sei X eine Menge an multinomial verteilten Zufallsvariablen, die die jeweilige Anzahl an Versuchen enthält, bei denen E_i eintritt. Sei ferner $F(X; n, p_{\mathbf{G}})$ die zugehörige Verteilungsfunktion dieses multinomial-verteilten Prozesses. Ein Ranking τ der Länge k mit $\tau_{\mathbf{G}}$ Anzahlen geschützter Kandidatinnen aller Gruppen $\mathbf{G}$, gegebenen Minimalanteilen $p_{\mathbf{G}} = (p_1, \ldots, p_i, \ldots, p_{|\mathbf{G}|})$ und Signifikanzniveau α erfüllt genau dann die* Fair Representation Condition*, wenn $F(\tau_{\mathbf{G}}; k, p_{\mathbf{G}}) > \alpha$.*

Definition 2 (Ranked Group Fairness Condition) *Ein Ranking τ erfüllt genau dann die* Ranked Group Fairness Condition *mit Parametern $p_{\mathbf{G}}$ und α, wenn für jedes Ranking-Präfix $\tau_{1..j}$ mit $1 \leq j \leq k$ gilt, dass es die Fair Representation Condition mit Minimalanteilen $p_{\mathbf{G}}$ und Signifikanzniveau α_c = AdjustSignificance$(k, p_{\mathbf{G}}, \alpha)$ erfüllt. (Die Funktion AdjustSignificance$(k, p_{\mathbf{G}}, \alpha)$ wird benötigt, weil wir mit dieser Vorgehensweise k* voneinander abhängige *Hypothesen testen. Sie gibt ein korrigiertes Signifikanzniveau α_c zurück, sodass die Gesamtsignifikanz der k Fairnesstests α beträgt).*

Mittels dieser Definitionen können wir nun eine Datenstruktur erzeugen, die für einen gegebenen Satz an Parametern $(k, p_{\mathbf{G}}, \alpha)$ die Minimalanzahl an Kandidatinnen jeder geschützten Gruppe an jeder Position enthält. Eine solche Struktur bezeichne ich als *mTree*. Abbildung 1 zeigt ein Beispiel eines mTrees mit den Parametern $p_{\mathbf{G}} = (0.2, 0.4)$ und $\alpha = 0.1$. Es sei angemerkt, dass jeder Pfad durch den mTree den Ranked Group Fairness

Test besteht und somit eine mögliche Verteilung von geschützten Kandidatinnen im fairen Ranking darstellt. Zur Erzeugung fairer Rankings orientiert sich FA*IR ebenfalls am jeweiligen mTree. Zunächst werden alle Kandidatinnen innerhalb ihrer jeweiligen Gruppe G_i nach Qualität sortiert. Anschließend ordnet FA*IR die Kandidatinnen in ein faires Ranking τ, indem immer die nächstbeste Kandidatin gewählt wird, außer mindestens eine Gruppe erfüllt ihre Minimalanzahl aus dem mTree noch nicht. Sollte dies der Fall sein, wählt FA*IR denjenigen Pfad, der laut $p_{\mathbf{G}}$ am wahrscheinlichsten ist (d.h. den höchsten Wert für F aufweist). Beispielsweise würde der Algorithmus für den mTree in Abbildung 1 an Position 4 den unteren Knoten wählen, da $p_1 < p_2$ und somit $F_{p_1} < F_{p_2}$. Sind zwei Pfade gleich wahrscheinlich, entscheidet FA*IR sich zufällig.

Fairness Dimensionen. FA*IR gehört zu den probabilistischen post-processing Methoden und arbeitet unter der Annahme, dass Entscheidungen fair sind, wenn Kandidatinnen mittels eines multinomialverteilten Zufallsprozesses (würfeln) positioniert werden. Dies bewirkt, dass die Relevanz der Kandidatinnen nur innerhalb ihrer Gruppen miteinander verglichen wird, nicht aber über Gruppenzugehörigkeit hinweg. Eine solche Annahme entspricht substanzieller Chancengleichheit und der Weltanschauung WAE (es sei jedoch angemerkt, dass der Algorithmus auch die Anwendung formeller Chancengleichheit und der WYSIWYG-Anschauung unterstützt, wenn die Minimalanteile $p_{\mathbf{G}}$ auf niedrige Werte gesetzt werden). FA*IR geht davon aus, dass alle messbaren Unterschiede zwischen den Gruppen in den Verteilungen ihrer Qualifikation durch vergangene Diskriminierung entstehen, die Verteilungen intrinsischer Relevanz (d.h. in *CS*) jedoch gleich sind. Der Algorithmus mindert demnach bestehenden und technischen Bias. Er kann mit mehreren Gruppen gleichzeitig umgehen und ist insbesondere für den Umgang mit intersektioneller Diskriminierung geeignet, da *alle* Gruppen *stets* voneinander getrennt betrachtet werden.

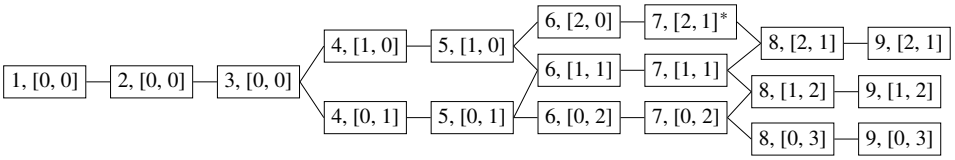


Abb. 1: Ein mTree für zwei geschützte Gruppen mit Minimalanteilen $p_{\mathbf{G}} = (0.2, 0.4)$ und bereits korrigiertem Signifikanzniveau α_c. Zu lesen ist die Notation $(k, [x, y])$ folgendermaßen: an der k-ten Position im Ranking müssen mindestens x Kandidatinnen aus Gruppe 1 und y Kandidatinnen aus Gruppe 2 gezählt worden sein, um Ranked Group Fairness zu erfüllen. Betrachten wir als Beispiel den markierten Knoten $(7, [2, 1])^*$: Damit ein Ranking als fair gemäß Definition 2 gilt, müssen in den ersten 7 Positionen mindestens 2 Kandidatinnen aus Gruppe 1 und 1 Kandidatin aus Gruppe 2 zu finden sein. Die ungeschützte Gruppe wird keiner Fairnessbedingung unterworfen, da sie keine Diskriminierung erfährt und demnach keiner besonderen Behandlung bedarf, um vergangene Ungerechtigkeit auszugleichen.

5 DELTR: Ein fairer Lern-Algorithmus für Rankings

Mit DELTR stelle ich den ersten fairen Learning-to-Rank (LTR) Algorithmus vor. Dieser erweitert die Verlustfunktion der bekannten Methode ListNet [Ca07] um einen Term, der die (Un-)Fairness $\mathcal{D}(.)$ des von Model f erzeugten Rankings misst. Somit werden Modelfehler L und Unfairness $\mathcal{D}$ gleichzeitig minimiert:

$$\underset{q \in \mathbf{Q}}{\text{minimize}} \quad L_{\text{fair}}\left(y^{(q)}, \hat{y}^{(q)}\right) = L\left(y^{(q)}, \hat{y}^{(q)}\right) + \gamma \cdot \mathcal{D}\left(\hat{y}^{(q)}\right)$$

wobei $y_i^{(q)}$ die Relevanz von Dokument d_i für Suchanfrage (auch Query) q darstellt. Die Relevanzen aller (Trainings-) Dokumente werden im Vektor $y^{(q)}$ zusammengefasst (analog stellt $\hat{y}^{(q)}$ den Vektor dar, der von f prognostiziert wird). Der Gewichtungsfaktor γ dient der Abwägung von Fairness versus Fehlerfreiheit. Für L verwendet ListNet die Kreuzentropie, ein Maß für die Verschiedenheit zweier Wahrscheinlichkeitsverteilungen.[5] Als Unfairnessmaß $\mathcal{D}$ verwende ich die Diskrepanz zwischen der *Exposition* zweier Gruppen in einem Ranking.

$$\mathcal{D}\left(\hat{y}^{(q)}\right) = \max\left(0, \text{Exposure}(G_0) - \text{Exposure}(G_1)\right)^2$$

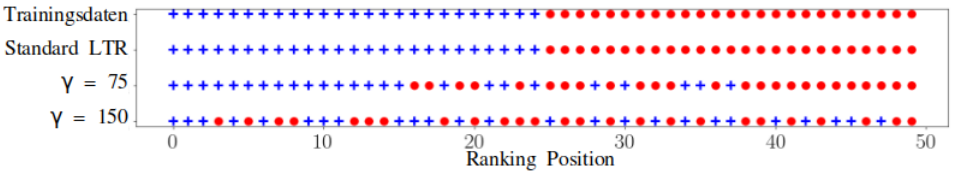
Dafür definiere ich zunächst die Exposition eines Dokuments d_i als die Wahrscheinlichkeit $P_{\hat{y}^{(q)}}$, dass d_i die erste Position im Ranking einnimmt. Die Exposition einer Gruppe G_a ergibt sich dann aus der durchschnittlichen Exposition jedes der Gruppe zugehörigen Dokuments:

$$\text{Exposure}(G_a) = \frac{1}{|G_a|} \sum_{d_i^{(q)} \in G_a} \text{Exposure}(d_i^{(q)})$$

Zu beachten ist, dass es sich bei $\mathcal{D}$ um ein asymmetrisches Maß handelt, das nur dann Bedeutung erlangt, wenn die geschützte Gruppe G_1 weniger exponiert ist, als die nicht-geschützte Gruppe G_0. Abbildung 2 verdeutlicht die Funktionsweise der Methode anhand eines kleinen synthetischen Datensatzes für verschiedene γ-Werte.

Fairness Dimensionen. DELTR gehört zu den expositionsbasierten in-processing Methoden, die entwickelt wurden um einige Nachteile der post-processing Methoden zu überwinden. Der entscheidenste Nachteil im Post-Processing besteht darin, dass ein unfaires Model trainiert wird, dass später korrigiert werden muss. Dies hat unter Wissenschaftlerinnen den Eindruck erweckt, dass Fairness und Genauigkeit stets gegeneinander abgewegt werden müssten. Jedoch zeigen meine Experimente mit DELTR, dass die Reduzierung von Bias zu einer verbesserten Genauigkeit des Models führen kann, wenn beide Zielfunktionen gleichzeitig optimiert werden. Die Methode mindert vor allem präexistierenden und technischen Bias. Da sie, wie auch schon FA*IR, kein Relevanzmaß in ihrer Fairnessdefinition enthält, ist sie der WAE-Anschauung und substanzieller Chancengleichheit zuzuordnen.

[5] Mittels einfacher mathematischer Kniffe, deren Erklärung ich aus Platzgründen auslasse, ist es möglich die Relevanzmaße aus $y^{(q)}$ und $\hat{y}^{(q)}$ in Wahrscheinlichkeiten umzuwandeln.



(a) In den Trainingsdaten befinden sich alle nicht-geschützten Dokumente auf den ersten Rankingpositionen.

(b) In den Trainingsdaten befinden sich alle geschützten Dokumente auf den ersten Rankingpositionen.

Abb. 2: Wirkung von DELTR anhand eines synthetischen Datensatzes. **Oben:** DELTR reduziert die Expositionsdiskrepanz für die geschützte Gruppe (rot). **Unten:** Durch die Asymmetrie verhält DELTR sich wie ein normaler LTR-Algorithmus, wenn die geschützte Gruppe bereits exponierter ist, als die nicht-geschützte Gruppe.

6 Zusammenfassung, Weiterführungen und Ausblick

Mit dieser Arbeit stelle ich verschiedene Möglichkeiten vor, gegen Diskriminierung in Rankings vorzugehen, wobei die Rankings sowohl für die Nutzerin relevant bleiben, als auch Fairness für die Kandidatinnen gegeben sein sollen. Dafür erarbeite ich zunächst ein breites Spektrum an philosophischen Grundlagen für Fairness und stellte deren Zusammenhang zu mathematischen Definition von Fairness her. Anschließend schlage ich zwei Methoden vor, die auf Prinzipien von substanzieller Chancengleichheit beruhen, einen In-processing, expositionsbasierten Ansatz und einen Post-processing, probabilistischen Ansatz. Bei beiden handelt es sich um die ersten ihrer Art.

Über diese Zusammenfassung hinaus enthält die Dissertation zu jeder Methode eine detaillierte Erarbeitung der mathematischen Theorie, ausführliche Experimente und deren Auswertungen, sowie Diskussionen zu den Grenzen der Ansätze. Zwei Kapitel lasse ich hier aus Platzgründen unberücksichtigt: Eines enthält eine umfangreiche Aufstellung anderer fairer Ranking-Methoden aus der Literatur, in der technische Funktionsweisen und damit verbundene normative Entscheidungen aufgezeigt und in Beziehung gesetzt werden. Das andere beschreibt die eingangs erwähnte Implementierung aller Algorithmen in FairSearch, die erste open-source API für faire Suche. Dies stellt die hier entwickelten Methoden einem breiten Publikum zur Verfügung, ohne die Notwendigkeit die gesamte wissenschaftliche Literatur zu durchforsten.

Literatur

[Ca07] Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th international conference on Machine learning. S. 129–136, 2007.

[Cr90] Crenshaw, K.: Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stan. L. Rev. 43/, S. 1241, 1990.

[FN96] Friedman, B.; Nissenbaum, H.: Bias in Computer Systems. ACM Trans. Inf. Syst. 14/3, S. 330–347, 1996, URL: https://doi.org/10.1145/230538.230561.

[FSV16] Friedler, S.; Scheidegger, C.; Venkatasubramanian, S.: On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236/, 2016.

[He19] Heidari, H.; Loi, M.; Gummadi, K. P.; Krause, A.: A moral framework for understanding fair ml through economic models of equality of opportunity. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, S. 181–190, 2019.

[JK02] Järvelin, K.; Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20/4, S. 422–446, 2002.

[Jo17] Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: ACM SIGIR Forum. Bd. 51, Acm New York, NY, USA, S. 4–11, 2017.

[ON17] O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books, 2017, ISBN: 0553418831.

[ZC20] Zehlike, M.; Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. In: Proceedings of The Web Conference 2020. S. 2849–2855, 2020.

[Ze17] Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; Baeza-Yates, R.: Fa*ir: A fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, S. 1569–1578, 2017.

[Ze22a] Zehlike, M.: Fairness in Rankings. Humboldt Universitaet zu Berlin (Germany), 2022.

[Ze22b] Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Sühr, T.; Baeza-Yates, R.: Fair Top-k Ranking with Multiple Protected Groups. Information Processing and Management 59 (1)/, 2022.

[ZSC20] Zehlike, M.; Sühr, T.; Castillo, C.: A Note on the Significance Adjustment for FA* IR with Two Protected Groups. arXiv preprint arXiv:2012.12795/, 2020.

[ZYS22a] Zehlike, M.; Yang, K.; Stoyanovich, J.: Fairness in ranking, part I: Score-based ranking. ACM Computing Surveys 55/6, S. 1–36, 2022.

[ZYS22b] Zehlike, M.; Yang, K.; Stoyanovich, J.: Fairness in ranking, part II: Learning-to-rank and recommender systems. ACM Computing Surveys 55/6, S. 1–41, 2022.

Meike Zehlike schloss ihre Dissertation 2022 mit dem Prädikat *Summa Cum Laude* im Fach Informatik an der Humboldt-Universität zu Berlin ab. Sie arbeitete zum Thema *Fairness und Diskriminierung in Suchalgorithmen* unter Prof. Ulf Leser, Prof. Carlos Castillo (UPF Barcelona) und Prof. Krishna Gummadi (Max-Planck-Institut für Software-Systeme). Ihre Dissertation setzt sich aus mehreren wegweisenden Veröffentlichungen zusammen, die zu den meistzitierten im Bereich Fairness in Rankings gehören. Für ihre Arbeit erhielt sie zahlreiche Auszeichnungen und Forschungsstipendien, unter anderem den Humboldt-Preis 2022 für die beste Dissertation, das Google-Stipendium Women Techmaker Award 2019, sowie den Data Transparency Lab Research Grant 2017. Sie ist seit 2019 Mitglied der ELLIS society (European League of Learning and Intelligent Systems).

Meike arbeitet heute als Senior Research Scientist bei Zalando, als externe Dozentin an der Humboldt-Universität, wo sie die Vorlesung “Trustworthy Machine Learning” hält, und als selbstständige Beraterin für die Gewerkschaft IG Metall und deren zugeordnete Betriebsräte. Neben ihren wissenschaftlichen Publikationen trägt sie viel zum Transfer ihrer Forschungsergebnisse in die Praxis bei: sie veröffentlicht regelmäßig Zeitungsartikel, tritt in Podcasts und Fernsehdokumentationen auf, und beteiligt sich an Normierungsinitiativen für KI-Systeme.