Herstellung von Transparenz für algorithmische Entscheidungen in Social Computing-Systemen

Bringing transparency to algorithmic decision making in social computing systems

Gummadi, Krishna P.

Max-Planck-Institut für Softwaresysteme, Standort Kaiserslautern, Kaiserslautern

Korrespondierender Autor

E-Mail: gummadi@mpi-sws.org

Zusammenfassung

Social Computing-Systeme sind eine sich entwickelnde Kategorie von Human-Computer-Systemen. Dazu zählen beispielsweise soziale Netzwerkseiten wie Facebook und Google Plus, Blogging- und Microblogging-Dienste wie Twitter oder LiveJournal, anonyme Social Media-Seiten wie Whisper und 4chan, Content Sharing-Plattformen wie YouTube und Instagram, Social Bookmarking-Websites wie Reddit oder Pinterest, Crowdsourcing-gestützte Meinungsseiten wie Yelp und eBay-Verkäuferratings und Social Peer Production-Plattformen wie Wikipedia und Amazons Mechanical.

Summary

Social computing systems refer to an emerging class of societal-scale human-computer systems. Examples include social networking sites like Facebook and Google Plus, blogging and microblogging sites like Twitter and LiveJournal, anonymous social media sites like Whisper and 4chan, content sharing sites like YouTube and Instagram, social bookmarking sites like Reddit and Pinterest, crowdsourced opinion sites like Yelp and eBay seller ratings, and social peer production sites like Wikipedia and Amazon's Mechanical Turk.

Von außen betrachtet werden Rechensysteme, auf denen die Sozialen Netzwerke basieren, durch die Interaktionen zwischen hunderten Millionen bis Milliarden von freiwilligen Nutzern angetrieben. In ihrem innersten Kern sind sie jedoch so gestaltet, dass sie die Handlungsweisen und Interaktionen ihrer Nutzer erfassen, vorhersagen und beeinflussen. Gesteuert werden sie über leistungsstarke Algorithmen, welche die Nutzerdaten erfassen und verarbeiten, um herauszufinden, welche Informationen die Nutzer auf Social-Media-Plattformen wie Facebook oder Twitter sehen, welchen Inhalten und Nutzern etwa auf YouTube und Wikipedia zu trauen ist und welche Produkte und Dienstleistungen zu welchem Preis auf eBay oder Yelp angeboten werden. Die Auswirkungen von Entscheidungen, die anhand solcher Mechanismen getroffen werden, können erheblich sein, da hunderte Millionen von Nutzern involviert sind. So ist beispielsweise bekannt, dass der Facebook-Algorithmus, der die angezeigte Reihenfolge von neuen Nachrichten für Facebook-Nutzer festlegt, einen wesentlichen Einfluss darauf hat, welche Meldungen auf beliebten Medienseiten wie NewYorkTimes.com oder CNN.com die größte Aufmerksamkeit erhalten.

Problem – mangelnde Transparenz

Hinsichtlich der Funktionsweise dieser Algorithmen mangelt es uns heute allerdings an Transparenz. So wissen wir insbesondere nur sehr wenig darüber, wie die Entscheidungsalgorithmen im Kern der Systeme arbeiten. Wir verstehen also nicht, welche Nutzerdaten in den Algorithmus einfließen, wie diese Systeme die Nutzerdaten verarbeiten, um ihre Entscheidungen zu treffen, und welche Auswirkungen die Entscheidungen auf ihre Nutzer haben. Angesichts der Tatsache, dass die Entwicklung und die Kontrolle dieser Systeme in der Hand einiger weniger Unternehmen liegen, wirft diese fehlende Transparenz einige grundlegende Fragen nach der Verwendung von privaten Nutzerdaten zur Entscheidungsfindung und nach dem Potenzial von einseitigen und diskriminierenden Entscheidungen auf.

Lösung – Überprüfung von algorithmischen Entscheidungsprozessen

Unsere aktuellen Forschungsarbeiten konzentrieren sich darauf, solche algorithmischen Entscheidungsprozesse transparent zu machen, indem wir sie von außerhalb überprüfen. Hierzu sammeln wir die Ausgaben (Output) des Algorithmus als Reaktion auf mehrere (oft kontrollierte) Eingaben (Input) und verfolgen, wie die Ergebnisse bei veränderten Eingaben variieren. Beispiel: Nehmen wir an, wir wollen herausfinden, ob sich bei Facebook der Standort eines Nutzers darauf auswirkt, in welcher Reihenfolge die Nachrichten seiner Freunde auf dessen Homepage angeordnet werden. Bei diesem Szenario erstellen wir ein neues Profil auf Facebook, das mit dem ersten Nutzerprofil identisch ist, das heißt dieselben Freunde und Interaktionen besitzt, mit Ausnahme des Feldes „Ort". Dann überprüfen wir, ob sich die Reihenfolge der Nachrichten in vorhersehbarer Weise ändert. Anders ausgedrückt: Um das Ausmaß zu überprüfen, in dem sich Informationen über einen Nutzer (also Eingaben in den Algorithmus) auf den Service auswirken, den dieser Nutzer erhält (also die Ausgabe des Algorithmus), verändern wir die Eingaben und beobachten die entsprechenden Output-Änderungen. Tatsächlich konnten wir aufgrund einer ausreichenden Anzahl von Ein- und Ausgaben maschinelle Lerntechniken anwenden, um den Algorithmus durch Reverse-Engineering zu rekonstruieren.

Eine Fallstudie – Coverage-Bias in den Suchergebnissen und Empfehlungen in Social Media

In unserer aktuellen Studie wollten wir verstehen, welche algorithmischen Entscheidungen hinter den Suchergebnissen und Empfehlungen beliebter Social-Media-Webseiten wie Facebook oder Twitter zu einer einseitigen Nachrichtenverbreitung führen. Da sich Twitter in den USA zu einer beliebten Informationsquelle für politische Nachrichten entwickelt hat, haben wir uns auf die Einseitigkeit von Twitter-Geschichten (Tweets) über Präsidentschaftskandidaten konzentriert. Anlass sind die aktuellen, parteiinternen Vorwahlen 2016, bei denen die Demokraten und die Republikaner in den USA ihre Kandidaten für die Präsidentschaftswahl aufstellen. Dabei wollten wir herausfinden, ob die Tweets, die die Nutzer mit der Twitter-Suchmaschine gefunden haben oder die den Nutzern von der Twitter-Maschine empfohlen wurden, eine verzerrte Berichterstattung ergeben haben. - Ob es also Geschichten waren, die von Nutzern mit einer ideologischen Tendenz zu Gunsten einer der Parteien (Demokraten oder Republikaner) gepostet wurden.

Bei den Suchergebnissen und Empfehlungen in den Sozialen Medien zu diesem Thema ergab sich hohe sogenannte Coverage-Biases, für die sich verschiedene Quellen (Ursachen) ausmachen ließen. Eine Ursache für diese Verzerrung sind die Eingaben, d. h. eine unausgewogene Altersstruktur der Nutzer, die über Social

Media Geschichten verbreitet haben. Im Vergleich zu ihrem tatsächlichen Anteil an der Gesamtgesellschaft äußert sich ein wesentlich größerer Anteil aller Social Media-Nutzer zugunsten der einen Partei (Demokratische Partei) als der anderen (Republikanischen Partei). Demzufolge steht ein erheblich höherer Anteil der über Suchergebnisse und Empfehlungen in Twitter gefundenen Geschichten ideologisch einer Partei näher.

Eine andere Quelle von Verzerrungen entsteht durch eine Personalisierung empfohlener Beiträge. Dabei werden Algorithmen dazu verwendet, Rückschlüsse darauf zu ziehen oder vorherzusagen, welche Geschichten für einen einzelnen Nutzer interessant oder uninteressant sein könnten, und ihm dann nur selektiv die interessant erscheinenden Beiträge präsentiert. Um diese Auswahl nutzergerecht maßzuschneidern, verwenden Social-Media-Plattformen wie Twitter verschiedene persönliche Informationen, die ihnen über den Nutzer zur Verfügung stehen. So empfiehlt Twitter einem Nutzer beispielsweise Geschichten, die Freunde des Nutzers (also Nachbarn des Nutzers im sozialen Netzwerk Twitter) interessant fanden. Solche Empfehlungen werfen wegen der Verwendung personenbezogener Informationen des Nutzers nicht nur Datenschutzfragen auf. Sie geben auch Anlass zur Sorge darüber, dass Nutzer sozusagen in „Filterblasen" gefangen sind, in denen sich stets Nutzer mit ähnlichen Ideologien bewegen. Es fehlt darin also an unterschiedlichen Perspektiven, die ihre Weltsicht in Frage stellen könnten. Durch diese einseitige Präsentation von Themen erhöhen solche Empfehlungsalgorithmen das Risiko einer gesellschaftlichen Polarisierung.

Interessanterweise lassen sich solche Suchergebnis- und Empfehlungsalgorithmen auch derart umgestalten, dass die bei unserer Analyse festgestellten Verzerrungen vermieden werden. Die technische Herausforderung besteht darin, zunächst Maßnahmen zur Bias-Quantifizierung vorzuschlagen und die Algorithmen dann so zu einzuschränken, dass Ausgaben (Outputs) erzeugt werden, die sich innerhalb akzeptabler Bias-Grenzwerte bewegen. Die Entwicklung von solchen fairen, nicht verzerrenden Algorithmen ist Gegenstand aktiver Forschungsarbeiten. Wir gehen davon aus, dass in den kommenden Jahren auf diesem Gebiet wichtige Fortschritte gemacht werden.

Zukünftige Richtungen – Jenseits von Social Computing-Systemen

Neben Computing-Systemen, die für online genutzte Sozialen Netzwerke arbeiten, wird die algorithmische Entscheidungsfindung auch zunehmend in verschiedenen Offline-Bereichen wie etwa im Bankgeschäft bei der Bonitätsbeurteilung, im Justizwesen zur vorausschauenden Überwachung und bei der Personalsuche für die Vorauswahl von Bewerbern verwendet. Dabei wird die automatisierte und datengestützte Entscheidungsfindung oft als Möglichkeit angepriesen, systematische Fehler und Ineffizienzen zu vermeiden, wie sie in großen Organisationen und Regierungen in der Vergangenheit aufgrund menschengemachter Entscheidungen immer wieder vorgekommen sind. Da die automatisierte Datenanalyse aber die menschliche Steuerung und Intuition in Entscheidungsprozessen ersetzt und der Umfang analysierter Daten „sehr groß" wird, wächst die Sorge über einen möglichen Verlust von Transparenz, Verantwortlichkeit und Fairness. So beruhen beispielsweise viele Algorithmen auf historischen Daten über von Menschen getroffene Entscheidungen. Kommen diese allerdings durch diskriminierendes Verhalten zu Stande (beispielsweise wenn ethnische Gruppen anderen Gruppen vorgezogen werden), setzt sich diese Verzerrung oder Diskriminierung in künftigen Entscheidungsprozessen, die von den Algorithmen durchgeführt werden, fort.

Deshalb ist es wichtig, dass wir nicht nur die von den Algorithmen getroffenen Entscheidungen überprüfen, sondern auch Methoden für eine Neugestaltung der Algorithmen vorschlagen, die potenzielle Verzerrungen in deren Entscheidungen eliminieren.

