Proposal zur Bachelorarbeit

...

von

Ricardo Valente de Matos

Matrikelnummer: 7203677
im Studiengang Wirtschaftsinformatik
der Fachhochschule Dortmund

Erstprüfer: Prof. Dr.-Ing. Guy Vollmer
Zweitprüfer: Stephan Schmeißer, M. Sc., Adessoplatz 1, 44269 Dortmund

Dortmund, den 12. Februar 2024

Motivation

In einer globalisierten und dynamischen Wirtschaftswelt sind Unternehmen zunehmend auf Projekte angewiesen, um ihre Ziele zu erreichen und Wettbewerbsvorteile zu erlangen. Die Personalbeschaffung für solche Projekte erfordert oft spezialisiertes Fachwissen und vielfältige Fähigkeiten, um erfolgreich umgesetzt zu werden. Es ist entscheidend für den Projekterfolg, dass die Personalbeschaffung die passenden Mitarbeiter für ausgewählte Projekte findet. Hier setzt die Entwicklung eines Recommender Systems zur Mitarbeiterempfehlung an. Ein solches System kann Unternehmen dabei unterstützen, den Prozess der Mitarbeiterrekrutierung und -auswahl zu optimieren. Durch die Berücksichtigung verschiedener Kriterien wie Qualifikationen, Fähigkeiten, Erfahrungen kann das Recommender-System dazu beitragen, die Auswahl effektiv zu filtern und diejenigen herauszufiltern, die am besten zu einem Projekt im Unternehmen passen. Ein solches System bietet außerdem den Vorteil, den Prozess der Mitarbeiterempfehlung zu automatisieren und zu beschleunigen. Dies ermöglicht Unternehmen, schneller auf offene Stellen zu reagieren und potenzielle Kandidaten zeitnah zu identifizieren. Dadurch wird die Effizienz der Mitarbeitersuche verbessert und die Qualität der Einstellungsentscheidungen erhöht.

Das Potenzial von Recommender Systems wurde auch bei *adesso* entdeckt und nun wird nach und nach Wege gesucht, KI-gestützte Systeme in die eigenen Prozesse zu integrieren. Im internen Projekt *adesso Staffing Advisor* wird an einem Recommender-System zur Mitarbeiterempfehlung für ausgewählte Projekte gearbeitet. Die Umsetzung der Recommender Systems bedient sich verschiedener KI-basierten Ansätze. Ein ganz entscheidender Schritt im Prozess der Mitarbeiterempfehlung ist die Vorverarbeitung der Bedarfsmeldungen. Diese sind eine wertvolle Informationsquelle, die Fachkräften helfen kann, die Empfehlungen effizienter zu gestalten, um dadurch wettbewerbsfähig zu bleiben. Allerdings sind diese oft umfangreich, unsortiert und komplex, was ihre effektive Nutzung erschwert.

Deshalb ist es entscheidend, effiziente Methoden und Techniken des Information Retrieval anzuwenden, um so relevante Informationen schnell und präzise aus Bedarfsmeldungen zu extrahieren. Die Extraktion wichtiger Schlüsselwörter, Phrasen und Themen ermöglicht es einen besseren Einblick in die Ziele, Methoden und Ergebnisse der Projekte zu bekommen. Dadurch können fundierte Entscheidungen bezüglich der Personalbesetzung getroffen und Ressourcen effizient genutzt werden.

Problemstellung

In einer immer stärker vernetzten und informationsreichen Welt stehen Organisationen vor der Herausforderung, relevante Informationen effizient aus umfangreichen Bedarfsmeldungen zu extrahieren. Obwohl diese Beschreibungen wichtige Einblicke in Ziele, Methoden und Ergebnisse liefern, können sie aufgrund ihres Umfangs und ihrer Komplexität schwer durchsuchbar und analysierbar sein. Die manuelle Identifizierung und Extraktion relevanter Inhalte ist zeitaufwendig und fehleranfällig. Daher stellt sich die Problemstellung:

Wie können wir effektive Methoden und Techniken des Information Retrieval und Data-Mining nutzen, um automatisiert relevante Inhalte aus Bedarfsmeldungen im spezifischen Software Entwicklungs-Kontext zu extrahieren und somit die Effizienz, Genauigkeit und Geschwindigkeit der Informationsgewinnung für Führungskräfte zu verbessern.

In der Vergangenheit wurden bereits Methoden im Bereich des automatisierten Recruitings untersucht. Im Projektgeschäft sehen wir uns mit einem Problem konfrontiert, dessen Umfang jedoch präziser definiert werden kann, da die Kandidatenauswahl einem begrenzten Pool unterliegt. Besondere Relevanz hat hierbei die Standardisierung der Bedarfsmeldung, da diese häufig unstrukturiert vorliegt und in variabler Ausgestaltung präsentiert wird.

Ziele und Ergebnisse der Arbeit

Diese Ausarbeitung präsentiert eine umfassende Untersuchung zur Entwicklung eines automatisierten Systems zur Extraktion relevanter Inhalte aus Bedarfsmeldungen im Software-Entwicklungs-Kontext.

- Die erste Phase dieser Ausarbeitung besteht darin, eine klare Erwartungshaltung hinsichtlich der Anforderungen und Bedürfnisse der Stakeholder zu entwickeln. Hierfür werden Interviews mit Führungskräften durchgeführt, um die Erwartungen bezüglich einer „perfekten“ Bedarfsmeldung herauszuarbeiten. Diese dient als Grundlage für die weiteren Entwicklungs- und Evaluierungsphasen.

- Im Anschluss erfolgt eine eingehende Analyse der Techniken <was für Techniken> des Information Retrieval und Data-Mining, um die besten Ansätze zur Extraktion relevanter Inhalte zu identifizieren. Diese Analyse bildet die Grundlage für die Konzeptionierung einer Vorverarbeitung, das eine Kombination der erforschten Ergebnisse darstellt. Die Implementierung dieses Modells erfolgt durch den Aufbau einer Pipeline in Python, die eine effiziente Verarbeitung und Extraktion der Bedarfsmeldungen ermöglicht.

- Zur Evaluierung der Leistungsfähigkeit des entwickelten Systems werden reale Bedarfsmeldungen und Mitarbeiterinformationen verwendet. Dabei wird überprüft, inwiefern das Ergebnis der definierten Erwartungshaltung entspricht. Mithilfe von den Metriken *Precision*, *Recall* und *F1-Score* werden Abweichungen, Ähnlichkeiten und Anpassungen in Parametern analysiert, um Erkenntnisse darüber zu gewinnen, wie das System inhaltlich abschneidet und verbessert werden kann.

- (Schließlich wird eine vergleichende Untersuchung mit einem auf Large Language Model basierenden Vorverarbeitungsansatz durchgeführt. Dabei werden die Performance, Zeit und Ergebnisqualität des entwickelten Systems mit diesem alternativen Ansatz verglichen. Dieser Vergleich dient dazu, die Stärken und Schwächen des entwickelten Systems zu identifizieren und gegebenenfalls weitere Verbesserungen vorzunehmen.)

Vorgehen und Zeitplan

Ziel ist es die Arbeit im Mai fertig zu stellen. Die einzelnen Monatsziele können aus der nachfolgenden Tabelle entnommen werden.

Februar	• Durchführung der Interviews mit Fachkräften • Zusammentragung aller relevanter Information Retrieval- und Preprocessing-Ansätze
März	• Durchführung der Interviews mit Fachkräften • Formulierung der Anforderungen für Bedarfsmeldungen
April	• Entwicklung des Eigenen Preprocessing-Modells • Evaluierung der Ergebnisse
Mai	• Schluss schreiben • Korrekturen

Aufbau der Arbeit

1 Einleitung **9**

1.1 Problemstellung 10
1.2 Ziele und Ergebnisse der Arbeit 11
1.3 Aufbau der Arbeit 11

2 Literaturüberblick **12**

2.1 Definitionen und Konzepte: Information Retrieval, Data-Mining, Bedarfsmeldungen 12
2.2 Vorherige Forschung und Ansätze zur Personalbeschaffung und Vorverarbeitung von Bedarfsmeldungen 12
2.3 Relevante Methoden und Techniken im Bereich Information Retrieval und Data-Mining 12

3 Entwicklung einer klaren Erwartungshaltung **13**

3.1 Beschreibung der Interviews mit Führungskräften zur Identifizierung von Stakeholder-Erwartungen 13
3.2 Analyse der Ergebnisse und Entwicklung einer klaren Erwartungshaltung für die Bedarfsmeldungen 13

4 Analyse der Techniken des Information Retrieval und Data-Mining **14**

4.1 Beschreibung der untersuchten Techniken und Ansätze 14
4.2 Bewertung und Auswahl der besten Ansätze für die Extraktion relevanter Inhalte aus Bedarfsmeldungen 14

5 Konzeptionierung und Implementierung der Vorverarbeitung **15**

5.1 Beschreibung des entwickelten Vorverarbeitungsmodells basierend auf den ausgewählten Techniken 15
5.2 Details zur Implementierung der Pipeline in Python für die effiziente Verarbeitung von Bedarfsmeldungen 15

6 Evaluierung des entwickelten Systems **16**

6.1 Beschreibung des verwendeten Datensatzes und der Evaluierungsmethodik 16
6.2 Präsentation und Diskussion der Ergebnisse basierend auf den Metriken Precision, Recall und F1-Score 16
6.3 Analyse von Abweichungen, Ähnlichkeiten und Verbesserungspotenzialen des Systems 16

7 Zusammenfassung und Ausblick **17**

[8]
[20]
[4]
[6]
[3]
information filtering [13]
preprocessing [1]
——- spam-filter [19] [7] [22] —- TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF ist eine statistische Methode, die verwendet wird, um die Relevanz eines Begriffs in einem Dokument relativ zu einem Korpus von Dokumenten zu bestimmen. Wörter mit höheren TF-IDF-Werten werden als potenzielle Schlüsselwörter betrachtet. [2] [18]
Text-Ranking-Algorithmen: Text-Ranking-Algorithmen wie TextRank oder YAKE (Yet Another Keyword Extractor) verwenden Graphen-basierte Methoden, um Schlüsselwörter in einem Text zu identifizieren. Diese Algorithmen bewerten die Wichtigkeit von Wörtern basierend auf ihrer Verbindung zu anderen Wörtern im Text und extrahieren Schlüsselwörter entsprechend ihrer Rangfolge. [14] [23] [16]
N-Gramm-Analyse: N-Gramme sind Sequenzen von N aufeinanderfolgenden Wörtern in einem Text. Durch die Analyse von N-Grammen können häufig auftretende Phrasen oder Begriffe identifiziert werden, die potenzielle Schlüsselwörter darstellen. [17]
Part-of-Speech (POS) Tagging: POS-Tagging wird verwendet, um die grammatischen Kategorien von Wörtern in einem Text zu bestimmen. Durch die Berücksichtigung von Wörtern mit bestimmten POS-Tags wie Substantiven oder Adjektiven können relevante Schlüsselwörter extrahiert werden. [12] [15]
bekommen wir ein unsupervised learning ansatz der
Regelbasierte Ansätze: Regelbasierte Ansätze verwenden vordefinierte Regeln oder Muster, um Schlüsselwörter zu identifizieren. Dies kann beispielsweise das Extrahieren von Wörtern sein, die häufig im Text vorkommen oder bestimmten Mustern entsprechen.
Hybride Ansätze: Hybride Ansätze kombinieren verschiedene Methoden und Techniken, um eine genauere Extraktion von Schlüsselwörtern zu ermöglichen. Zum Beispiel könnte eine Kombination aus TF-IDF-Gewichtung und Text-Ranking-Algorithmen verwendet werden, um eine robuste Schlüsselwortextraktion zu erreichen.
transformation von bedarfsmeldung zu guter bedarfsmeldung, was ist der fokus von der bedarfsmeldung, wie gut machen die ansätze das, und muss man das dann noch weiter verarbeiten, haben wir alles was wir brauchen mit nur einem algorithmus, inferenz falls parameter fehlt, gibt es einen der alles löst,

ergebnis der arbeit: diese modelle in der reihenfolge kommen am nähesten an die bedarfsmeldung
ausblick: die keyword extraction auch für die profile nutzen
was muss ich jetzt machen: gucken wie ich das inhaltlich genau machen will, also pipeline genauch checken, quellen von der bachelorarbeit checken

1 Einleitung

1.1 Problemstellung

1.2 Ziele und Ergebnisse der Arbeit

1.3 Aufbau der Arbeit

2 Literaturüberblick

2.1 Definitionen und Konzepte: Information Retrieval, Data-Mining, Bedarfsmeldungen

2.2 Vorherige Forschung und Ansätze zur Personalbeschaffung und Vorverarbeitung von Bedarfsmeldungen

2.3 Relevante Methoden und Techniken im Bereich Information Retrieval und Data-Mining

3 Entwicklung einer klaren Erwartungshaltung

3.1 Beschreibung der Interviews mit Führungskräften zur Identifizierung von Stakeholder-Erwartungen

3.2 Analyse der Ergebnisse und Entwicklung einer klaren Erwartungshaltung für die Bedarfsmeldungen

4 Analyse der Techniken des Information Retrieval und Data-Mining

4.1 Beschreibung der untersuchten Techniken und Ansätze

4.2 Bewertung und Auswahl der besten Ansätze für die Extraktion relevanter Inhalte aus Bedarfsmeldungen

5 Konzeptionierung und Implementierung der Vorverarbeitung

5.1 Beschreibung des entwickelten Vorverarbeitungsmodells basierend auf den ausgewählten Techniken

5.2 Details zur Implementierung der Pipeline in Python für die effiziente Verarbeitung von Bedarfsmeldungen

6 Evaluierung des entwickelten Systems

6.1 Beschreibung des verwendeten Datensatzes und der Evaluierungsmethodik

6.2 Präsentation und Diskussion der Ergebnisse basierend auf den Metriken Precision, Recall und F1-Score

6.3 Analyse von Abweichungen, Ähnlichkeiten und Verbesserungspotenzialen des Systems

7 Zusammenfassung und Ausblick

Ausblick

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt sowie die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Dortmund, den 12. Februar 2024

_____

Ricardo Valente de Matos

Literatur

[1] S. A. Alasadi und W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, Jg. 12, Nr. 16, S. 4102–4107, 2017.

[2] P. Bafna, D. Pramod und A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, 2016, S. 61–66.

[3] N. J. Belkin und W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" *Communications of the ACM*, Jg. 35, Nr. 12, S. 29–38, 1992.

[4] W. B. Croft, "Combining approaches to information retrieval," in *Advances in Information Retrieval: Recent Research from the center for intelligent information retrieval*, Springer, 2000, S. 1–36.

[5] L. De Angelis, F. Baglivo, G. Arzilli u. a., "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health," *Frontiers in Public Health*, Jg. 11, S. 1 166 120, 2023.

[6] R. Horesh, K. R. Varshney und J. Yi, "Information retrieval, fusion, completion, and clustering for employee expertise estimation," in *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, 2016, S. 1385–1393.

[7] A. Khorsi, "An overview of content-based spam filtering techniques," *Informatica*, Jg. 31, Nr. 3, 2007.

[8] M. Kobayashi und K. Takeda, "Information retrieval on the web," *ACM computing surveys (CSUR)*, Jg. 32, Nr. 2, S. 144–173, 2000.

[9] E. Kommission. "A european approach to artificial intelligence." (o. J.), Adresse: `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence` (besucht am 3. Jan. 2024).

[10] E. Kommission. "Commission welcomes political agreement on artificial intelligence act*." (April 2021), Adresse: `https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473` (besucht am 3. Jan. 2024).

[11] E. Kommission. “Der Ansatz der EU für künstliche Intelligenz konzentriert sich auf Exzellenz und Vertrauen, um Forschung und industrielle Kapazitäten zu stärken und gleichzeitig Sicherheit und Grundrechte zu gewährleisten.” (o. J.), Adresse: `https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence` (besucht am 3. Jan. 2024).

[12] D. Kumawat und V. Jain, “POS tagging approaches: A comparison,” *International Journal of Computer Applications*, Jg. 118, Nr. 6, 2015.

[13] C. Lanquillon, “Enhancing text classification to improve information filtering,” Diss., Otto-von-Guericke-Universität Magdeburg, Universitätsbibliothek, 2001.

[14] R. Mihalcea und P. Tarau, “Textrank: Bringing order into text,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, S. 404–411.

[15] T. Nakagawa und K. Uchimoto, “A hybrid approach to word segmentation and pos tagging,” in *Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 2007, S. 217–220.

[16] T. Pay, S. Lucci und J. L. Cox, “An ensemble of automatic keyword extractors: TextRank, RAKE and TAKE,” *Computación y Sistemas*, Jg. 23, Nr. 3, S. 703–710, 2019.

[17] S. Pirk, “Implementierung und Visualisierung N-Gramm-basierter Word-Clouds,” B.S. thesis, 2019.

[18] J. Ramos u. a., “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, Citeseer, Bd. 242, 2003, S. 29–48.

[19] M. A. Shafi’I, M. S. Abd Latiff, H. Chiroma u. a., “A review on mobile SMS spam filtering techniques,” *IEEE Access*, Jg. 5, S. 15 650–15 666, 2017.

[20] A. Singhal u. a., “Modern information retrieval: A brief overview,” *IEEE Data Eng. Bull.*, Jg. 24, Nr. 4, S. 35–43, 2001.

[21] S. Stowasser, O. Suchy, N. Huchler u. a., “Einführung von KI-Systemen in Unternehmen,” *Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München*, 2020.

[22] K. Tretyakov, “Machine learning techniques in spam filtering,” in *Data Mining Problem-oriented Seminar, MTAT*, Citeseer, Bd. 3, 2004, S. 60–79.

[23] M. Zhang, X. Li, S. Yue und L. Yang, "An empirical study of TextRank for keyword extraction," *IEEE access*, Jg. 8, S. 178 849–178 858, 2020.