جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Latifa RASSAM

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le vendredi 25 octobre 2024 à 14h30 au Grand Amphi à l'ENSIAS de Rabat

Intitulé de la thèse

Contribution to Textual Document Indexing: A Novel Fuzzy Logic-based N-gram Graph Indexing Approach

Président :

Pr. Soumia ZITI, PES, Faculté des Sciences, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Ahmed ZELLOU, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Sanaa EL FILALI, PES, Faculté des Sciences Ben M'Sick, Université Hassan II, Casablanca

Pr. Youssouf EL ALLIOUI, MCH, Faculté Polydisciplinaire-Khouribga, Université Sultan Moulay Slimane, Beni Mellal

Pr. Khalid HOUSNI, MCH, Faculté des Sciences, Université Ibn Tofail, Kénitra

Examinateur:

Pr. Samir ANTER, MCH, FST-Mohammedia, Université Hassan II, Casablanca

Résumé: L'imprécision dans l'indexation des documents textuels constitue un défi majeur dans le domaine du Traitement Automatique du Langage Naturel (TALN), entraînant souvent une réduction de la précision dans les tâches de récupération d'information et de fouille de données. L'ambiguïté et la variabilité inhérentes au langage nécessitent des techniques robustes pour gérer et atténuer efficacement ces problèmes d'imprécision, qui sont essentiels pour améliorer la fiabilité des systèmes de TALN.

Le graphe n-gramme, un outil puissant dans le domaine du TALN, joue un rôle crucial dans la capture des relations entre les mots et les phrases au sein d'un document. En représentant le texte sous forme d'un graphe d'interconnexions de n-grammes, cette approche permet une analyse plus structurée et complète des données textuelles. Cependant, les méthodes traditionnelles de graphes n-grammes peinent souvent à gérer l'imprécision, notamment dans les textes complexes et ambigus.

La logique floue, avec sa capacité à gérer l'incertitude et le raisonnement approximatif, offre une solution prometteuse aux problèmes d'imprécision dans l'indexation. En intégrant la logique floue dans le processus d'indexation, il devient possible de mieux capturer les nuances et les subtilités du langage, conduisant à des représentations du contenu textuel plus précises et fiables.

La première contribution majeure de cette thèse est le développement d'un nouvel algorithme d'indexation par graphe n-gramme, spécifiquement conçu pour améliorer la précision et l'efficacité de la représentation textuelle. Cet algorithme exploite les forces des graphes n-grammes tout en répondant aux limitations des approches traditionnelles, en particulier pour gérer les textes ambigus et imprécis.

En outre, cette thèse introduit deux autres contributions : la proposition de nouvelles fonctions basées sur la logique floue pour la génération d'index, et le développement de nouvelles fonctions basées sur la logique floue pour l'extraction des noms composés. Ces contributions visent à résoudre le problème de l'imprécision en fournissant des mécanismes d'indexation plus précis et sensibles au contexte, essentiels pour améliorer les applications de TALN.

Les méthodes proposées ont été rigoureusement testées sur des ensembles de données réels, démontrant que le nouvel algorithme et les fonctions basées sur la logique floue offrent d'excellentes performances en pratique. Les résultats montrent une amélioration significative de la précision, soulignant l'efficacité des approches proposées pour traiter les problèmes d'imprécision dans l'indexation des documents textuels.

Mots-clés: Graphe N-gram, Traitement Automatique du Langage Naturel, Indexation de Documents Textuels, Indexation Floue.

Abstract: Imprecision in textual document indexing poses a significant challenge in the field of Natural Language Processing (NLP), often leading to reduced accuracy in information retrieval and data mining tasks. The inherent ambiguity and variability in the language require robust techniques to effectively manage and mitigate these imprecision issues, which are critical for improving the reliability of NLP systems.

The n-gram graph, a powerful tool in the NLP domain, is crucial in capturing the relationships between words and phrases within a document. By representing text as a graph of interconnected n-grams, this approach allows for a more structured and comprehensive analysis of textual data. However, traditional n-gram graph methods often need help with handling imprecision, particularly in complex and ambiguous text.

Fuzzy logic, with its ability to handle uncertainty and approximate reasoning, offers a promising solution to the imprecision issues in indexing. By incorporating fuzzy logic into the indexing process, it becomes possible to better capture the nuances and subtleties of language, leading to more accurate and reliable representations of textual content.

The first major contribution of this thesis is the development of a novel algorithm for n-gram graph indexing, specifically designed to enhance the precision and effectiveness of text representation. This algorithm leverages the strengths of n-gram graphs while addressing the limitations of traditional approaches, particularly in handling ambiguous and imprecise text.

In addition, the thesis introduces two further contributions: the proposal of novel fuzzy logic-based functions for index generation and the development of new fuzzy logic-based functions for extracting compound nouns. These contributions are aimed at resolving the imprecision problem by providing more accurate and context-aware indexing mechanisms.

The proposed methods have been rigorously tested on real-world datasets, demonstrating that both the novel algorithm and the fuzzy logic-based functions perform exceptionally well in practice. The results show a significant improvement in accuracy, highlighting the effectiveness of the proposed approaches in addressing imprecision issues in textual document indexing.

Keywords: Fuzzy indexing, natural language processing, N-gram graph, textual document indexing.