Sistem Tanya Jawab menggunakan Knowledge Graph mengenai Sistem Tata Surya

Jaish Muhammad[1], Kemas Rahmat Saleh Wiharja, S.T., M.Eng, P.hD[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]jaishm@student.telkomuniversity.ac.id,
[2]bagindokemas@telkomuniversity.ac.id,

Abstrak

Manfaat dari *Knowledge Graph* (KG) bisa kita amati langsung, seperti optimalisasi search engine query Google, yang mempermudah dalam mencari sesuatu di internet. Diatas KG bisa juga dibangun sebuah *Question Answering System* (QA). Penelitian ini menggunakan artikel-artikel mengenai sistem tata surya dari halaman web NASA. Dengan menggunakan NLTK, artikel-artikel yang didapatkan dari halaman web NASA dipecah ke dalam bentuk *triple*. *Triple* tersebut kemudian diubah ke dalam bentuk *Knowledge Graph* yang disimpan di dalam Neo4j, kemudian dibangun *QA System* diatasnya. Proses validasi hasil sistem melibatkan ahli di bidang Astronomi. Hasil dari sistem ini adalah sistem yang menjawab query pertanyaan mengenai sistem tata surya. Performansi sistem diukur menggunakan akurasi, precision, recall, f1 score, dan mean reciprocal rank, yang mana didapatkan Akurasi = 0.78, Precision = 0.5, Recall = 1, F1 Score = 0.67, dan MRR = 0.2112955621.

Kata kunci : *Knowledge Graph*, *Question Answering system*, NASA, *triple*, sistem tata surya.

Abstract

The benefits of the Knowledge Graph (KG) can be observed directly, like the optimization of search engine queries like Google. which makes it easier to search for something on the internet. Based on the KG, a Question Answering System (QA) can also be built. This research uses articles about the solar system from NASA's website. Using NLTK, the articles obtained from NASA's website are broken down into triples. These triples are then transformed into a Knowledge Graph stored in Neo4j, and a QA System is built on top of it. The validation process of the system's results involves experts in the field of Astronomy. The outcome of this system is a system that answers queries about the solar system. The system's performance is measured using accuracy, precision, recall, F1 score, and mean reciprocal rank, which are obtained as follows: Accuracy = 0.78, Precision = 0.5, Recall = 1, F1 Score = 0.67, dan MRR = 0.2112955621.

Keywords: Knowledge Graph, Question Answering system, NASA, triples, solar system.

1. Pendahuluan

Latar Belakang

Sampai saat ini sudah cukup banyak penelitian yang menggunakan *Knowledge Graph* untuk *Question Answering System* [1][2][3][4][5][6][7][8][9][10]. Secara umum, *Knowledge Graph* menggambarkan suatu objek/entitas dan hubungan antara mereka [23]. *Knowledge graph* merupakan sebuah graf data yang dimaksudkan untuk mengumpulkan dan menyampaikan pengetahuan tentang dunia nyata, di mana nodenya mewakili suatu entitas dan edge mewakili hubungan yang berbeda antara entitas tersebut [24].

Fungsionalitas dari *Knowledge Graph* itu tidak terbatas, mulai dari answering search query di search engine, perbankan, retail, industri otomotif, industri perminyakan, kesehatan dan farmasi, penerbitan dan media [13]. Manfaat yang didapat dari *Knowledge Graph* antara lain lebih dari sekadar search, dapat memperoleh insight secara otomatis, rekomendasi secara otomatis, hingga analisis prediktif [14].

Implementasi dari *Question Answering system* tidak dibangun menggunakan *Knowledge Graph* saja, terdapat beberapa penelitian yang menggunakan Knowledge Base sebagai dasarnya. Sebuah KB-QA system mengambil natural language utterance sebagai input dan menghasilkan satu atau lebih jawaban sebagai output [15][16][17][18].

Berdasarkan studi literatur yang dilakukan, artikel-artikel yang didapatkan dari website NASA dapat diekstraksi dan dipecah ke dalam bentuk *triple*. *Triple* yang berhasil diekstraksi disimpan dalam bentuk *Knowledge Graph* menggunakan Neo4j. Kemudian, *Question Answering System* dibangun di atas *database* yang berhasil dibuat. Dataset yang digunakan berdasarkan halaman web Nasa yang memuat artikel mengenai sistem tata surya.

Topik dan Batasannya

Penelitian ini memiliki rumusan masalah sebagai berikut:

1. Bagaimana cara untuk membangun *Knowledge Graph* dari sebuah artikel?
2. Bagaimana cara untuk membangun *Question Answering system* menggunakan *Knowledge Graph*?

Batasan masalah:

1. Pengerjaan tugas akhir ini hanya sampai berhasil mengeluarkan jawaban dari pertanyaan yang diajukan.
2. Dataset dibatasi hanya dari satu sumber saja yaitu NASA, tidak terlalu luas.

Tujuan

Tujuan dari pelaksanaan tugas akhir ini adalah:

1. Mengetahui cara membangun sebuah *Knowledge Graph* dari sebuah artikel.
2. Mengetahui cara membangun *Question Answering system* menggunakan *Knowledge Graph*.
3. Berhasil mendapatkan jawaban yang relevan dari *Question Answering System* yang dibangun.

Organisasi Tulisan

Bagian lanjutan dari penelitian ini akan diisi dengan bagian 2 yaitu studi terkait, yang berisi studi-studi yang mendukung penelitian yang dilakukan. Selanjutnya pada bagian 3 merupakan sistem yang dibangun berdasarkan rancangannya. Pada bagian 4 dilakukan evaluasi dari sistem yang dibangun, terdapat penjelasan mengenai hasil pengujian sistem yang dibangun. Bagian 5 merupakan kesimpulan yang didapat dari hasil penelitian yang dilakukan, juga berisi saran yang bisa dilakukan untuk penelitian lanjutan.

2. Studi Terkait

2.1. Solar System

Solar System atau Tata Surya, merupakan kumpulan benda langit yang terdiri atas sebuah bintang yang disebut Matahari dan semua objek yang terikat oleh gaya gravitasinya [22]. Anggota utama dari tata surya terdiri dari Surya/Matahari, Merkurius, Venus, Bumi, Mars, Jupiter, Saturnus, Uranus, dan Neptunus[22]. Penelitian ini menggunakan artikel-artikel mengenai tata surya yang didapat dari laman web Nasa.

2.2. Knowledge Graph

Sebuah *knowledge graph* (i) sebagian besar menggambarkan entitas dunia nyata dan keterkaitannya, terorganisir dalam bentuk graf, (ii) mendefinisikan kelas dan relasi dari entitas yang memungkinkan dalam sebuah skema, (iii) memungkinkan untuk potensi *interrelating arbitrary entities* antar satu sama lain dan (iv) mencakup berbagai domain topik [12]. Pada penelitian ini, *Knowledge Graph* dibangun berdasarkan artikel-artikel dari halaman web Nasa, dan disimpan di dalam Neo4j.

2.3. NLTK

NLTK (Natural Language Toolkit), adalah sebuah rangkaian modul program sumber terbuka, tutorial, dan set masalah, yang menyediakan perangkat pengajaran linguistik komputasional yang siap digunakan [21]. Pada penelitian ini, NLTK digunakan sebagai library utama untuk proses ekstraksi *Triple* yang selanjutnya diubah ke dalam bentuk *Knowledge Graph*.

2.4. spaCy

spaCy adalah *library* Python gratis dan sumber terbuka yang menyediakan kemampuan canggih untuk melakukan pemrosesan bahasa alami (NLP) pada volume teks yang besar dengan kecepatan tinggi [25].

2.5. Named Entity Recognition (NER)

Named Entity Recognition (NER) merupakan metode pemrosesan bahasa alami (NLP) yang berfungsi untuk mendapatkan informasi dari sebuah kalimat. NER merujuk pada subjek utama dari sebuah teks, seperti nama, lokasi, perusahaan, acara dan produk, serta tema, topik, waktu, nilai moneter dan persentase [38].

2.6. Neo4j

Neo4j merupakan sebuah *database*, tetapi berbeda dengan *database* pada umumnya yang menyimpan data dalam baris, kolom dan tabel, Neo4j memiliki struktur yang fleksibel. Neo4j menyimpan relasi yang menghubungkan data-data. Dengan Neo4j, setiap data record atau node mengandung direct pointers ke seluruh node yang terkoneksi [19].