Hans Sandbu

Spatiotemporal Keyword search on RDF graphs

Master's Thesis in Computer Science, Spring 2020

Norwegian University of Science and Technology

Abstract

This thesis looks at methods for using keywords for search in RDF data, with the goal of finding what aspects are needed for accuracy and speed. To accomplish this, some previous methods for keyword searching will be recreated, and then extended to incorporate both spatial and temporal searches.

Sammendrag

Husk at hvis du er en norsk student og skriver masteren din på engelsk, så *må* du lage et sammendrag på norsk.

(If you are a non-Norwegian student, it is not obligatory to include an abstract in Norwegian.)

Preface

Spatiotemporal keyword search on RDF graphs.

Hans Sandbu

Trondheim, 19th May 2020

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 2
- 1.2 Goals and Research Questions . . . 3
- 1.3 Research Method . . . 3

2 Related Work **5**
- 2.1 RDF graphs . . . 5
- 2.2 Keyword search on graphs . . . 5
- 2.3 Spatial search on RDF graphs . . . 6
- 2.4 Temporal search on RDF graphs . . . 7

3 Background Theory **9**
- 3.1 Graph theory . . . 9
  - 3.1.1 Graph traversal . . . 9
- 3.2 RDF, ontologies, and knowledge graphs . . . 10
  - 3.2.1 Ontology . . . 10
  - 3.2.2 RDF as knowledge graphs . . . 11
  - 3.2.3 Existing knowledge graphs and ontologies . . . 12
    - Yago . . . 12
  - 3.2.4 Uses for Knowledge graphs . . . 12
  - 3.2.5 Jena . . . 13

4 Architecture **15**
- 4.1 Search . . . 15
- 4.2 Ranking . . . 17
- 4.3 Pruning . . . 17
  - 4.3.1 Predicate selection . . . 18
  - 4.3.2 Place hierarchy . . . 19
  - 4.3.3 Bounding . . . 19

5 Experiments and Results **21**
- 5.1 Experimental Plan . . . 21
  - 5.1.1 Time . . . 21
  - 5.1.2 Scoring and ranking . . . 21
- 5.2 Experimental Setup . . . 22
  - 5.2.1 Data set . . . 22

5.2.2 Queries . . . 23
5.2.3 Combining data and queries . . . 23
5.2.4 Max. distance and following edges . . . 23
5.3 Experimental Results . . . 23
5.3.1 Effect of distance from root . . . 23
5.3.2 Effect of following specific edges . . . 24
5.3.3 Differences between data types . . . 24
5.3.4 Implications . . . 25

6 Discussion and limitations **29**
6.1 Discussion . . . 29
6.2 limitations . . . 31

7 Conclusion and Future Work **33**
7.1 Contributions . . . 33
7.2 Future Work . . . 33

Bibliography **35**

Appendices **37**

List of Figures

1.1 Example of non-directed graph. 1
1.2 Simple Directed graph. 1
1.3 Directed graph with a cycle. 2

3.1 Trondheim as a subject in YAGO 13
3.2 Connections in Yago around Elvis 14

5.1 Linear regression for time and roots 26
5.2 Time used for data types and traversal criteria 26

List of Tables

5.1 Platform used for experiments 22
5.2 Follow all edges, max. distance 2 24
5.3 Follow all edges, max. depth 1 25
5.4 Follow only edges from node hits 27

1 Introduction

In computer science, graphs are structures used for representing relationships between objects. A graph consists of vertices, sometimes called nodes, connected by edges, where the vertices represent the objects, and the edges represents the relationship. A graph can take on different shapes, giving the graph special properties. By giving the edges a direction the graph can take on further properties.

A graph where the edges have a direction is called a directed graph as seen in figure 1.2. Here the vertices are connected to each other, with an arrow displaying the direction. An example of an undirected graph can be seen in figure 1.1. In both types of graphs a vertex can have multiple edges. Graphs can contain cycles, meaning that vertices are connected to each other so that it is possible to follow edges in such a way that it leads back to the initial vertex. An example of such a cyclic graph can be seen in figure 1.3. Acyclic graphs have an inherent topological ordering. This ordering makes it possible to find properties in the graph, such as the shortest path between two vertices.

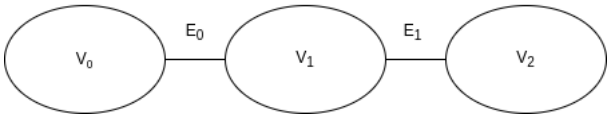

Figure 1.1: Example of non-directed graph.

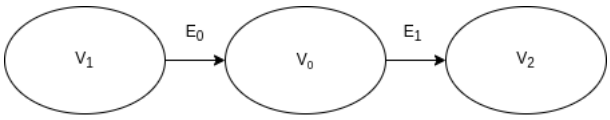

Figure 1.2: Simple Directed graph.

When using graphs as a form of storage, the vertices hold data, and the edges can be given an extra property describing the relation. Combining such a data store with common traversal methods for graphs results in an efficient extraction model for relational data. Some common traversal methods include breadth first search, where all neighboring vertices are discovered before moving on to the children of the currently discovered vertices. This contrasts to depth first search, where the children of a vertex is visited as they are discovered.

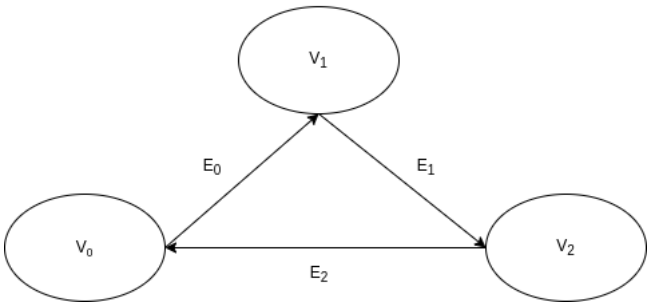

Figure 1.3: Directed graph with a cycle.

Databases using graphs have been created to overcome some of the limitations in traditional relational databases. Comparing such graph storage to relational storage, graph storage generally outperforms relational storage on structural queries, but not on data queries [20]. structural queries reference the graph structure, or the relationships in the data, while data queries look for specific data. When creating a graph database, an indexing system can also added. Such a system is "Neo4j"[1] that use Apache Lucene[2] for indexing.

Another model for storing graph data is the Resource Description Framework (RDF). RDF often have its own form of storage, a "triplestore" based on the structure of the data. In RDF data is stored as triples, where a triple have the structure "subject-predicate-object". In this structure the subject and object can be considered vertices, and the predicate is the edge. In addition to forming the relation between the two vertices, the predicate also holds a type of relation. This makes it possible to apply reason to the data set, and derive new information from what is already there.

There are a few projects using RDF as an information storage system. Some of the more well known are DBPedia[4], YAGO[17], and Creative Commons. These applications often consists of a large linked data set, or in the case of Creative Commons, it is used for embedding licenses. Creating methods that allows for easy access to information in RDF data makes it possible to create applications with utility. In this thesis methods for efficient keyword search on spatial and temporal data will be explored.

1.1 Background and Motivation

UMost of the information on the web is unstructured, or semi-structured data. Using metadata and bots to help create structures can help both humans and computers. This is one of the goals of RDF. Creating ways for humans to discover data in new ways can create utility. To be able to get mor utility out of RDF, making it more accessible to humans needs to be done. Because of the highly linked nature of data that exists now,

[1] https://neo4j.com/
[2] https://lucene.apache.org/

using these links can help provide that utility. As more data is generated, there are also more relations between the data. Exploring these relations can be done by using RDF or other graphs, but for a regular person this can be difficult. By proving that fast and accurate keyword search of spatiotemporal RDF graphs is possible new utilities for exploration can be built.

1.2 Goals and Research Questions

Goal *This thesis has the goal of researching methods for spatiotemporal keyword searches on large RDF graphs, and what aspects needs to be considered for speed and accuracy.*

Accomplishing this goal proves that RDF graphs can be used as a tool for structuring data related to real world places. There is a lot of data that can be placed at one or more real locations, either directly, or indirectly. By using RDF graphs it is possible to model how different places are connected through some pice of data, and how different pieces data can be related to real places.

Research question 1 *How can spatiotemporal data be integrated into exiting keyword query methods for RDF data?*

Extending existing methods for searching RDF data can make it simpler to search spatiotemporal data. By combining methods for temporal and spatial search, methods for spatiotemporal search should be possible to research.

Research question 2 *What methods can be used to achieve greater speed and accuracy for searches on RDF data?*

An important aspect of any search is speed and accuracy. Since searching in graphs usually entails some form of traversal, the methods of traversal, as well as factors that affect speed and accuracy will be researched.

Research question 3 *How do spatial and temporal RDF query methods differ from from other query methods?*

Comparing spatial, temporal and spatiotemporal searches can tell how spatiotemporal data differs in RDF graphs. This is important so that optimized methods can be developed.

1.3 Research Method

This thesis will build on previous keyword search approaches for RDF graphs, and determine what methods can be expanded to incorporate spatiotemporal queries. A method for spatiotemporal search will be created, and methods for improvement will be tested, with the goal of finding what aspects of the search method have most effect for speed and accuracy.

2 Related Work

Keyword search on RDF graphs, and methods for traversal have been researched before. This research forms the basis when extending search to include spatiotemporal data.

2.1 RDF graphs

In 1999 the World Wide Web Consortium (W3C) introduced the "Resource Description Framework Model and SyntaxSpecification" [2]. Here the first definitions of RDF was described. This was an XML based syntax designed to provide interoperability between applications on the web, in a machine readable format. By providing information in a machine readable fashion, the creation of automated processes should be easier to create, and by using a common standard, the same automated processes could read any page containing RDF data.

The data model of RDF can be compared to object oriented data. RDF consists of objects, literals, and the connection between them [5]. The objects can have literals connected to them, indicating some form of data, and the objects can be connected to each other. Connections are called predicates, and form the relation between an object and literal data, or form the relation between two objects. This forms a graph structure, where the objects and literals are vertices, and the predicates are edges.

Modeling RDF as a graph creates a directed graph [13]. In this model the predicates defines a direction between the two objects. Such a model is called a triple, consisting of a "Subject predicate object structure"[5]. Here the subject is the start vertex of the direction, and the objec denotes the end vertex.

2.2 Keyword search on graphs

Keyword search on RDF graphs often follows a set of common strategies, that usually involves graph traversal. One method is to find nodes containing one or more keyword, then following the edges from the nodes to explore the graph, and find subgraphs where the combined nodes contain as many keywords as possible while also spreading as little as possible. BLINKS [10] propose such a method, in combination with indexing and cost balancing for expanding clusters of accessed nodes. A similar approach is used by the authors in [7] where each node has an associated document containing terms from the

triple. When querying the keywords are matched with these documents, creating lists based on the matching keywords, and a subgraph is constructed by joining matches from different lists.

Another strategy for keyword search in RDF graphs is to infer triples from the query. One such query system is used in AquaLog [11]. This method processes the input into a triple based representation, based on a linguistic model, and then further processes the triplets into what they call "query triples." Creating structured queries through inference is also done in [19]. Here the query is first used to find nodes containing some part of the query, then the graph is explored to find a connection between the nodes. The result is a series of subgraphs connecting nodes that contain part of the query. Each of the subgraphs are in turn used to create a conjunctive query with edges mapped to predicates, and nodes to subjects or objects.

Ranking and scoring the results of a search is also needed for evaluating the different methods and algorithms. A common element for ranking the results is to look at the span of the subgraphs or trees returned from the search. The shorter distance between all nodes, the more accurate a result should be. Of the above mentioned papers, three [10, 7, 19] use some form of minimum spanning tree or graph when scoring or ranking the results. In addition to the minimum spanning graph, the results can be ranked by other factors.

BLINKS adds a scoring system where shared vertices are counted multiple time, once for each node connected to it. This is done to score trees with nodes close to the root higher than nodes further away, even if the further nodes have many shared edges. The content of the nodes are also scored based on an IR style TF/IDF method. In the paper by [7] the minimum tree are ranked by a probabilistic model, and a language model. The probabilistic model scores a result based on the average probability for a term to occur in a triple in the subgraph. In addition, the language model is used to score some keywords higher based on what part of the triple they are found in. This triple scoring is done by weighting words based on the structure of the triples they are found in, so keywords that occur more often in predicates are scored higher if they are found in a predicate. The final paper, [19], adds popularity, and keyword matching to the minimum spanning graph. Popularity is calculated based on how many edges a node has, so that the more connected a node is, the less cost a path through that node has. The keyword matching score is based on keyword matches in a node, but is also weighted based on syntactic and semantic similarity, which is in turn done by using WordNet data.

2.3 Spatial search on RDF graphs

Keyword searching and ranking spatial RDF graphs is quite similar to regular RDF keyword searching. [16] outlines methods that incorporates spatial data into the search. These methods is similar to some of those previously mentioned here [19, 7]. The most substantial difference is the use of R-trees to index the spatial dimension of the graph.

This is done so that a subgraph can be rooted both at a real point in the world, and nodes in the graph close to the real world point. The root is used as a point for traversing the graph, and like the previous methods, the goal is to find a minimum subgraphs. The subgraphs are ranked based on how close the root node is to the selected point, the size of the tree, and how well the result tree fits the query words.

2.4 Temporal search on RDF graphs

Using the syntax specification for RDF [1] it is possible to declare dates as literals. Using such literals, dates and times can be connected to any vertex using a user defined predicate. This makes it easy to add time and date data to any graph, but some problems arise [18]. When adding time with user defined predicates, semantics can be lost. A predicate such as "borneOnDate" denotes a start date for a human, but cannot be used as a start date for other concepts. It is possible to remedy this by attaching a concept such as "startDate" to the predicate, but this has the drawback of complicating the structure. Adding time as literals also means that the specific time is only connected to one vertex. This is a problem if a subgraph is temporal. It is a possibility to add predicates from all vertices in the subgraph to the date, but this has the drawback of creating many extra predicates.

Using literals for temporal data can be called "time labeling". Another method for adding time to RDF graphs is using snapshots, called "versioning". This method maintains the state of the graph at a given time. Such a model will create multiple versions of the graph to be able to model time. A versioned graph will be more difficult to traverse and query than one using literals [9].

Temporal data can come in two different forms, time points, and time intervals. Time points can be modeled using a single literal. Combining two literal time points, "$[a, b]$" can create a time interval, where "$a \leq b$". Here the nodes a and b is usually defined by the xsd datatype *date* [18]

Using query languages such as SPARQL it is possible to query time labels. In temporal queries, time labels can be queried directly, and filtered for use in a time interval query.

3 Background Theory

3.1 Graph theory

Graph theory is the study of graphs as mathematical structures that models relationships between objects in a collection. A graph consists of vertices connected by edges, *G(V,E)*. In computer science, graph theory is used in areas such as data mining, image segmentation, networking and more[15].

3.1.1 Graph traversal

When traversing a graph, two methods are commonly used, breadth first nd depth first. Both traversal methods needs a starting point. If there is no natural starting point or root, any vertex in the graph can be selected. Breadth first traversal use a queue containing vertices in the order they are discovered. From the start vertex, all connected vertices are added do the queue, then the first vertex in the queue is visited, and explored. Vertexes connected to the current node is added to the back of the queue, so that the oldest discovered node is the next to be visited. In contrast a depth first search will add newly discovered vertices to the front, so that the newly discovered vertex will be visited next.

Both traversal methods can be used on directed and undirected graphs. On an undirected graph, any vertex discovered can be added to the list of vertices to visit. In a directed graph, only vertices connected by an edge with a direction going from the current vertex, and to another vertex will be followed. Both traversal types can have a termination condition, so that the traversal will stop before traversing the entire graph.

Algorithm 1: BFS(Root R)

Result: BFS
Queue **Q** ADD(R);
Tree t ;
while **Q** $\neq \emptyset$ **do**
    e = POP(**Q**) ;
    n = GetNeighbors(e) ;
    Q ADD(n) ;
end

Algorithm 2: BFS(Root R)

```
Result: BFS
Stack S ADD(R);
Tree t ;
while S ≠ ∅ do
    v = POP(Q) ;
    if v not discovered then
        v.setDiscoveered ;
        n = GetNeighbors(v) ;
        Q ADD(n) ;
    end
end
```

Algorithm 3: MinimumSpanningTreeBFS(Root R, Terms Q_t)

```
Result: Minimum tree containgin terms Q_t
Queue Q ADD(R);
Tree t ;
MatchedWords M_w ;
while Q ≠ ∅ do
    e = POP(Q) ;
    if e.tokens ⊆ Q_t then
        t ADD(e) ;
        M_w.ADD(e.token ∩Q_t) ;
    end
    if Q_t = M_t then
        Break ;
    end
    n = GetNeighbors(e) ;
    Q ADD(n) ;
end
```

3.2 RDF, ontologies, and knowledge graphs

Knowledge base, ontology and knowledge graph are terms with multiple definitions, and are often used interchangeably. The definitions here are used to clarify what they mean in this paper, and is not a definitive definition.

3.2.1 Ontology

Ontologies contain representations, conceptualization, relations, categorization, and formal naming of data [3]. ontologies allows for semantic modeling of knowledge. Here

knowledge is data that is not ordered in a strict structure. Using ontologies allows data to be better structured, and reason about the semi-structured data.

Structuring an ontology using a framework like RDF creates a knowledge graph. There is no clear definition on exactly what constitutes a knowledge graph. A broad definition of knowledge graph is "A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge." [6] This definition is encompasses multiple technologies. Another definition using RDF graphs as a bases is "We define a Knowledge Graph as an RDF graph. An RDF graph consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject s $\in$ U $\cup$ B, a predicate p $\in$ U, and an object U $\cup$ B L. An RDF term is either a URI u $\in$ U, a blank node b $\in$ B, or a literal l $\in$ L" [8]. This latter definition is how knowledge is structured in the data used in this thesis.

3.2.2 RDF as knowledge graphs

Using RDF as a format when modeling knowledge makes it possible to create structured content from semi-structured knowledge. This structure makes knowledge usable for computer, where it previously would have been difficult to extract accurate and exact knowledge. This information can then be presented in a human readable fashion, or it can be used in other fields, like artificial intelligence.

RDF data is organized into triples. Each triple consists of a subject, object, and a predicate. In an RDF triple the subject is a URI or a blank node. The URI when used in the subject identifies an entity, or is an alias, different language or other variation of an entity. This subject can have relations to objects describing the same entity.

Predicates are always a URI. URIs used for predicates differ from the ones used for subjects and objects in that predicates are of a type. A type is an identifier used to describe the relation between the subject and object.

Objects have the widest range of possible entries. Like subjects and predicates, objects can be URIs. Object URIs can be entity identifiers like subjects, they can be class identifiers, or they can contain some data and a data type. Blank nodes are just that, blank, and literals are an atomic value.

A fact is a term often used when describing knowledge bases, ontologies and knowledge graphs. Usually a fact is the smallest piece of information in such a system. In a knowledge graph this is a single RFD triple. Another term often used is entity. An entity is a collection of facts, usually from the same article. Entities can be linked together through facts. In such a fact, one entity is used as the subject, the predicate describes the relation, and the object is the other entity. In such a relation both entities will be URIs.

3.2.3 Existing knowledge graphs and ontologies

Currently two of the largest open technologies for knowledge graphs are Yago and DBPedia. Both these projects use automatic extraction from Wikipedia to create the graph, but differ in the ontology used to build the graphs. Yago also includes data from WordNet and GeoNames to accurately assign entities to classes. Both projects use RDF triples to create a knowledge graph.

Yago

Yago is an acronym for Yet Another Great Ontology, and is main data source in this paper. The project describes it self as a knowledge base[17] and an ontology [12], but is often described as a knowledge graph by others. In this paper Yago is described as a knowledge graph.

Yago have many entities with facts describing date and spatial data.[17] Date facts follows the ISO 8601 format, YYYY-MM-DD, and introduces # as a wildcard symbol. A fact can only hold information on a single point in time, and uses yagoDate as a data type in addition to information on the date.[17] In a date fact, the object holds the date information, and the predicate describes a connection between subject and date. "Nidaros_Cathedral wasCreatedOnDate 1300-##-## ." In this example the predicate wasCreatedOn is used to describe a relation between the subject and a date.

To describe a time span two facts are required. One of the facts describes a start date, and the second en end date. Since an entity can have multiple date facts connected, all date predicates are also assigned to a class. Start dates are assigned to a predicate with a type that has a "creation" class, such as "StartedOnDate". End dates are assigned to "destruction" type predicates. This makes it possible to deduce a time span for a given subject and predicate combination.[17]

Yago only contains permanent spatial data for entities on earth. This means that entities like cities, buildings, rivers and mountains are given a spatial dimension. In addition, events, people, groups and artifacts can be given a spatial dimension by relating the entity to a specific place. All spatial facts must have a predicate that fall under the yagoGeoEntity class, and all objects used in a fact with a yagoGeoEntity must have a relation containing both "hasLatitude" and "hasLongitude".

3.2.4 Uses for Knowledge graphs

One of the uses of knowledge graphs today is to find and display a info box in search engines. This information is a compact set of facts that tries to fit the search query. Because of the graph structure of knowledge graphs the information in the info box can be adapted to the query by choosing the predicates and related facts closest related to the query. This makes it possible to create a set of information that can give the user a quick overview of the information retrieved by the query.

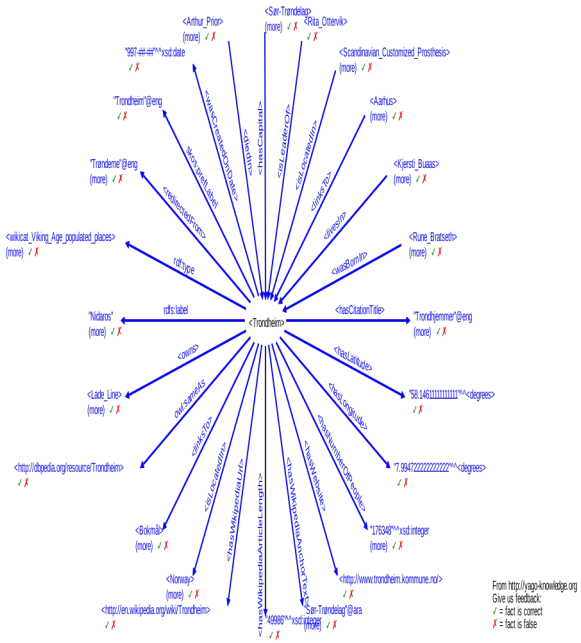

Figure 3.1: Trondheim as a subject in YAGO

3.2.5 Jena

Apache Jena is a framework for working with semantic web, and linked data like RDF. The framework contains tools for SPARQL querying, a query language made specifically for RDF graphs. Jena also contains REST-style SPARQL endpoints making the RDF data easily accessible. Using the existing standards makes it easy to use existing data sets, such as Yago or DBPedia, and build utility on top of that data using some of the tools Jena provides.

Jena provides a persistent storage solution called TDB[1]. This storage contains a table for vertices, and indexing of triples, and a prefix table. The node table stores the representation of RDF "terms", where terms are any vertex, excluding some literals. Excluded literals are the xsd:decimal, xsd:integer, xsd:dateTime, xsd:date, and xsd:boolean types. Each vertex is stored with an ID used when data is loaded, and when querying constant terms. Node to ID mappings are stored as a B+ tree, while the ID to node mappings are stored as a sequential access file. Triples are stored in the triple table as a tuple with three node IDs. The prefix table is used for supporting mappings that again are used mainly for serialization of triples. Queries in Jena are run using ARQ, a SPARQL supported query engine.

[1] https://jena.apache.org/documentation/tdb/architecture.html

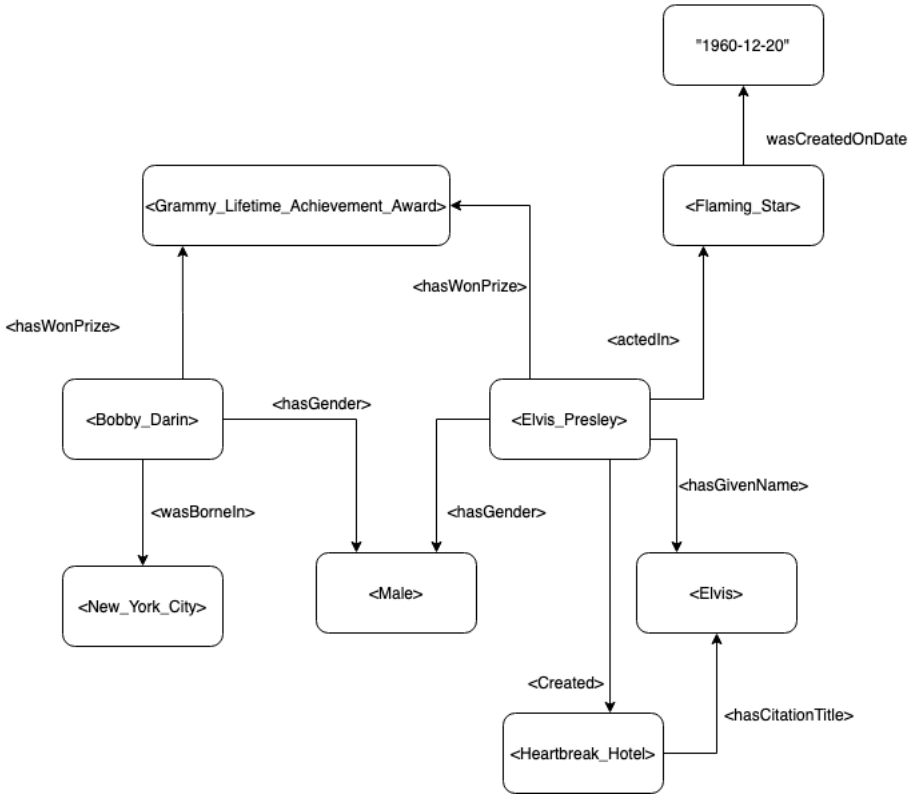

Figure 3.2: Connections in Yago around Elvis

4 Architecture

When building a system for keyword search, the first implementations was based on previously created systems. This early method was based on [16] and the method was also used as the base for adding a temporal dimension to the search.

4.1 Search

Each search will return a subgraph, based in a root. This forms a tree structure. This result can described as follows:

Result tree *Result r for a given query q with tokens Qt on an RDF graph G⟨V, E⟩, where the result forms a minimum spanning tree Tm⟨V', E'⟩, V' contains a set of tokens Tt, Tm is rooted in a place vertex p, so that V' ⊆ V, E'⊆E, and Tt⊆Qt*

All results from a query is given as a minimum spanning tree. This tree is call a *Result Tree*. A query can have multiple result trees, where each result tree is rooted in a unique place node, and each tree contains at least one query word.

A root node can be described as:

Root *Starting point of a BFS traversal, and the vertex all other vertices are connect to in a result tree*

A spatiotemporal root is like any other root, but with an addition:

spatiotemporal root *Any root R where the spatial input I_s and temporal input I_t overlaps, so that $R \in I_s \cap I_t$*

One basic method of finding a match for keywords is using a breadth first search (BFS) [10, 16]. For each keyword in the query the algorithm will find all vertices that contain the keyword. From that set of vertices the BFS search finds the first vertex that can connect all the vertices in the set. The vertex that connects the the rest of the set is the one that best fits the keywords in the search.

Using BFS starting on keyword vertices works for keyword search, but not necessarily when searching for spatial or temporal data. When adding spatial or temporal vertices to the search, these can be used as a root, and a starting point for traversal. When a search initiates, a place and/ or time vertex existing in the graph needs to be selected. From this vertex or vertices, all connected places will be used as roots when searching. For

the selected place vertex, the roots is geographically located within the selected place. This is done to bound the search close to the selected place. From the root nodes the search will traverse the graph looking for vertices matching the query words, and when all words are found, that subgraph is the best match for the start root. After all roots are searched, the subgraphs are given a score based on how well they fit the query.

The first part of the search is to find a set of root nodes for sub-graphs. This is done by collecting all neighbors from the place queried. The neighbors can be connected by any predicate in the first developed algorithm, but as explain in 4.3 this selecting specific predicates can greatly increase speed by limiting the amount of nodes traversed. The BFS algorithm described above is used for each of the possible roots of subgraphs, stopping when all keywords are matched, or the depth of the graph exceeds the three or exceeds the currently shallowest subgraph that has matched all keywords. The reason for limiting the depth of subgraphs is to ensure at least a partial hit within a reasonable time. Limiting the depth of subgraphs to the current shallowest subgraph is done because no a deeper graph cannot be a more accurate hit.

When traversing the graph and finding a match, a node object is created. This object is used to keep track of the pseudo hierarchy in the graph. All objects contain information on the depth of the node, matched query terms, and relation to parent and children, if any. In addition root objects contains a list of all query terms hit in the sub graph. If a child node is found within multiple subgraphs, a new object will be created representing the node for each subgraph. This creates some objects that are nearly identical, but with different relations and possible different depth.

When a node object is created, a document with tokens is created. This document is used to calculate score, and to match a node with the query words. A document is created from the last part of a Yago URI, after the last slash. Each URI si further split on each underscore, creating a list of words.

After traversing the main graph we have found all subgraphs containing at least one query term. These subgraphs contain many nodes which hit the same terms. Before ranking the subgraphs the minimum spanning graph needs to be found. The minimum spanning tree is found using a greedy algorithm that iterates through all nodes in the subgraphs, then keeping the nodes containing the most terms, and lowest depth. The minimum tree will contain as many terms as possible, a term will only be found in one node, and the graph will be as shallow as possible.

Graph traversal implementation

Algorithm 4: GetFullResultTree(p, t, Qt)

```
Result: Set containing all nodes with at least one keyword
Queue Q ADD(p);
Threshold t;
Set n ;
while Q ≠ ∅ do
    Clear(n) e=POP(Q) if GetDistance(e) > t then
        continue;
    end
    foreach Term t ∈ Qt do
        if t ∈ e then
            p.AddMatchChild(e);
        end
    end
    if Qt ⊆ e then
        t = GetDistance(e);
    end
    n = GetNeighbors(e) ;
    Q ADD(n);
end
```

4.2 Ranking

All result trees are given a score based on how well they fit the query. This score is called accuracy and is made up of two parts. The first part is how well a vertex fits the query words, and the other is how far it is removed from the root.

Accuracy: *Score given to a result tree, where score is based on the number of nodes n in the result tree, node distance nd from root, query q, and number of query words hit h so that* $\mathcal{F}_h = (h_i/\,|q| : i \in \mathcal{I})$ *and* $\frac{\sum \mathcal{F}_h}{n*(nd+1)}$

To find the accuracy, a minimum spanning tree is first extracted from the result tree. To extract this tree, the algorithm 5 is used.

4.3 Pruning

When using the BFS search method, all possible spatial vertices close to the queried place will be explored. This is expensive and many of the vertices will be irrelevant. Pruning the potential place vertices will reduce the amount of subgraphs traversed, and will in turn reduce the overall time used to find results for the query.

Algorithm 5: FindMinimumTree(R_t)

Result: A minimum spanning tree based on keyword vertices found during traversal

```
ResultTree R_t MinimumTree T_m ADD(R_t[0]);
foreach Vertex r ∈ R_t do
    foreach Vertex m ∈ T_m do
        if r.hits = m.hits ∧ r.distance > m.distance then
            Break;
        end
        if r.distance = m.distance ∧ r.hits = m.hits then
            Continue;
        end
        if r.hits ≠ m.hits then
            m.match REMOVE_COMMON_WORDS(r);
            T_m ADD(r);
        end
        if m.hits = ∅ then
            T_m REMOVE(m);
        end
    end
end
```

4.3.1 Predicate selection

When traversing the graph a lot of unnecessary predicates are followed, resulting in many extra nodes added to the search. Specifying a set of predicates that contain the relevant information can greatly increase the speed of the algorithm, and also keep the memory requirements a lot lower. When selecting predicates for traversal the information expected from the search should be the top priority. Because of this, all predicates that may contain spatial or temporal data should be kept.

When using the entirety of the Yago data set, most nodes are highly connected. Many of the links in the graph are from predicates such as “linksTo” or “redirectedFrom”. These predicates creates a highly connected graph, and ensures a hit within a few nodes of the start. The same predicates will also often add the same nodes multiple times, create circular graphs, and take up unnecessary CPU power and memory.

When pruning predicates there are two methods that are possible to implement. The first will remove the predicates that contain little or no new information, such as “linksTo” or “redirectedFrom” mentioned above. This will still keep the graph connected, and keeps the predicates containing more useful information.

An other method of pruning is to create a list of predicates to be followed. This can drastically reduce the connections in the graph, but the results will only contain information relevant to the query. When preselecting predicates there is a much greater

chance of not finding a match for a query. In addition a lot of metadata could be lost, if the metadata predicates are not added to the list of predicate to be explored.

4.3.2 Place hierarchy

For any spatial search, a lot of places will be found. The graph also contains information on the relationship between places. To find the best possible matches for a spatial query, only places of a higher resolution should be chosen. This means that places of similar or smaller expanse should be allows to be queried, e.g. a query of Boston can return the entirety of Boston, close to Boston, or within Boston. Massachusetts has multiple predicates linking it to Boston, but should not be queried because the information returned would be less detailed than what is queried.

4.3.3 Bounding

When selecting places or times for a query, the boundaries should be set so that they encompass the entirety of the queried time and or place.

5 Experiments and Results

5.1 Experimental Plan

The goal of the experiments is to gather data for comparison of data types searched and criteria for vertices to follow. when comparing, speed and accuracy will be the metrics used. In total, 9 different searches will be conducted, one using spatial vertices as a starting point, one for temporal vertices as start, and one combining the two, searching spatiotemporal vertices. For each of these vertex types, there will also be different criteria for how the graph will be explored. The first method follows all predicates, excluding connections to types and classes, and will have a max depth of 2. The second method follows the same predicates as the first, but has the max depth set to 1. Finally, a criteria for following only predicates from nodes with at least one keyword match is used. When comparing the methods, the goal is to see how reducing the amount of data search will affect speed and accuracy of the search, and how differences in depth and edges followed differs based on the starting data.

5.1.1 Time

When timing the search, two different times will be measured, time for each query, and time for each root node. In addition to these, avg. nodes visited for each root will also be used. Taking the average of these over a large input set should generate an appropriate result. Both time metrics measure how fast results are found, so the results should be similar as long as the input data does not contain a high number of highly connected nodes.

The timing of the algorithm should be the same as any breath first search, $\Theta(\mathsf{V} + \mathsf{E})$ so that the best case is visiting only the first vertex, and the worst case is traversing the entire graph.

5.1.2 Scoring and ranking

Each subgraph is given a score. This score will be used to see how pruning, and difference in predicates can lead to differences in the result tree. Two metrics for accuracy will be used, avg. accuracy for each result and avg. highest accuracy for each query.

5.2 Experimental Setup

All experiments was run on a single laptop with the specifications listed in table 5.1. The code is written in Java, using openJDK 11.0.7 and building with gradle 6.0. For triple store, Jena tdb storage was used, using version 3.14.0 of Jena.

OS	Ubuntu 19.10 x86_64
Host	20KGS0N400 ThinkPad X1 Carbon 6th
Kernel	5.3.0-46-generic
CPU	Intel i5-8350U (8) @ 3.600GHz
Memory	15760MiB

Table 5.1: Platform used for experiments

5.2.1 Data set

All of the data used is from YAGO. YAGO was selected for the large open data set, rich taxonomy, and for the spatial and temporal parts of the ontology. There are several data sets available for download, divided into categories. For running experiments data from taxonomy, core, and geonames were selected. From the taxonomy category, all data sets where used. This data describes class structure, entities, and defines relationships.

From the core category all data was also used. The core contains is most of the data used in the graph. This includes dates, relationships between nodes, literals, and labels. Most of the vertices and edges used in the experiment comes from this category, but this data can be further structured using some of the data from other categories.

Geonames contains data and structure for geographical vertices. These vertices have a hierarchical structure based on what places are located within others. In addition, the data contains literals for coordinates, alternative names and links for neighbors. In addition the category contains additional classes and types specific for the geographical vertex.

Before the data could be loaded into a store, some preprocessing was necessary. This includes replacing non-unicode character, and replacing spaces with underscore in URIs. In addition the data from YAGO contained som illegal characters for Jena, such as double quoted URIs and illegal escape sequences. There were also some unterminated ttl lines. All the data was run through a sed script to ensure correctly formatted data for Jena. After formatting the data correctly a persistent TDB storage was created using tdbloader from Jena.

5.2.2 Queries

When selecting words for the queries, all nodes in the graph was scanned to count word frequency. This list was sorted based on number of occurrences, stop words and numbers were removed. From this shortened list, the top 150 words were chosen, based on vocabulary size needed to meet more than half of nodes [14, 21]. From this list of 150 words, a final set of 10 words was chosen at random. It is worth noting that all words are from the name of a vertex, so the list of words does not reflect natural language.

When selecting places, a set of 7 places were manually chosen. This choice was made to ensure places from different parts of the world, and difference in population and language. With these differences roots found should have variation in edges, differences in keyword vertices discovered, and paths discovered for shared keyword vertices.

Date selection for temporal search was done manually. This selection included a combination of non-significant dates, and significant dates. The selection was made to have a high probability of finding hits, and ensure that the vertices have variation in the amount of edges.

5.2.3 Combining data and queries

When running the experiments, the same set of query words was used for all runs. For each type of data, spatial, temporal, and spatiotemporal, the query words was combined in five pairs of two, and four pairs of four. variation in the amount of words used would affect both the chance of hitting keyword vertices, and the score of the subgraphs discovered.

5.2.4 Max. distance and following edges

For each of the data types a total of three different exploration methods was used. One wit a max. distance of 1 from the root node, one with a max. distance of 2, and one where only edges leading to a keyword vertex would be explored further. Using a max. depth of 2 would serve as a base line for the comparison. The two other methods was used to see how much distance from the root would affect subgraphs accuracy, and if optimal graphs could be found even with a severe limitation put on the search.

5.3 Experimental Results

5.3.1 Effect of distance from root

When traversing a graph, the amount of nodes visited will heavily impact the time used. When the max. distance of the search increase, the amount of vertices visited will grow exponentially. Looking at table 5.2 we see that the amount of roots found for a query is the most significant factor for the speed and amount of nodes visited. Since there only is

three data points, no broad conclusion can be drawn, but in 5.1 a trend can be seen, and all points line up well. This indicates a linear increase of vertices visited, based on the amount of edges the vertices have.

Table 5.2: Follow all edges, max. distance 2

Traversal following all edges max. distance 2			
Result type	Temporal data	Spatial data	Spatial and temporal data
Avg. time per query (ms)	27 679	166 479	1 842
Avg. roots per query	192	1 527	10
Avg. nodes per root	3 097	3 684	4 843
Avg. nodes visited per query	594 683	5 626 989	49 719
Avg. accuracy per result	0.075	0.077	0.066
Avg. top accuracy per query	0.231	0.280	0.108
Query miss (no keywords found)	24/63 (38.1%)	6/63 (10.5%)	357/441 (81.0%)

5.3.2 Effect of following specific edges

In table 5.4 we see that very few vertices with a keyword are directly connected to other vertices with a keyword. This makes the search results similar to the the results in table 5.3. The most noticeable difference is the amount of vertices visited. Even with a greater max. depth, the search only following edges from vertices with keyword hits still visits fewer vertices. Visiting fewer vertices increase the speed of the search, something that is evident from graph 5.2.

5.3.3 Differences between data types

In table 5.3 we can see how many connections the average root have. Because the max. distance is set to 1, the average vertices per root is the same as average connections per root. Temporal data have the highest number of edges, followed by spatiotemporal vertices. This indicates that vertices that contain temporal data generally are more connected that spatial vertices. The difference in roots discovered can bne attributed to the input data.

Table 5.3: Follow all edges, max. depth 1

Traversal following all edges max. distance 1			
Result type	Temporal data	Spatial data	Spatial and temporal data
Avg. time per query (ms)	4 111	14 618	467
Avg. roots per query	192	1 527	10
Avg. vertices per root	450	257	335
Avg. vertices visited per query	86 479	391 914	3 446
Avg. accuracy per result	0.231	0.232	0.174
Avg. top accuracy per query	0.313	0.349	0.215
Query miss (no keywords found)	35/63 (55.5%)	20/63 (31.7%)	415/441 (94.1%)

5.3.4 Implications

Two factors are significant for the speed and accuracy of a search. The first is the set of roots used when searching. Limiting the amount of root nodes will increase the speed of the search. This in addition to decreasing the amount of edges followed have the most impact on speed. The other factor is the depth of the search. Increasing the depth will not guarantee greater accuracy on the top hits, but when comparing the miss rate of searches with distance one and two, a greater distance will reduce the miss rate.

Figure 5.1: Linear regression for time and roots

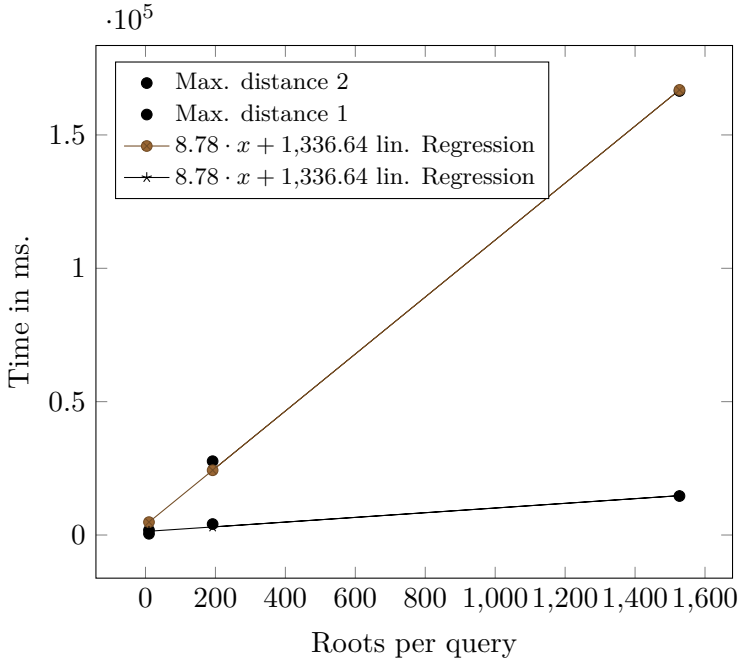

Figure 5.2: Time used for data types and traversal criteria

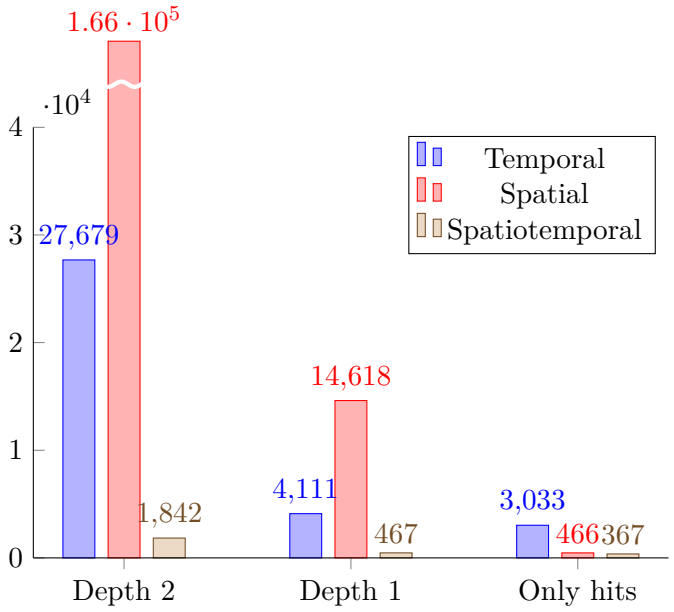

Table 5.4: Follow only edges from node hits

Traversal following edges with keyword match vertex			
Result type	Temporal data	Spatial data	Spatial and temporal data
Avg. time per query (ms)	3033	466	367
Avg. roots per query	192	1527	10
Avg. vertices per root	1.02	1.03	1.02
Avg. vertices visited per query	197	1569	11
Avg. accuracy per result	0.231	0.232	0.174
Avg. top accuracy per query	0.313	0.349	0.215
Query miss (no keywords found)	35/63 (55.5%)	20/63 (31.7%)	415/441 (94.1%)

6 Discussion and limitations

6.1 Discussion

Using traditional methods for graph traversal it is possible to implement keyword search on RDF graphs. One such method is the use of BFS to find a minimum spanning subgraph containing the query words. Using spatial nodes as a root for the subgraph, creating a minimum spanning tree, it is possible to retrieve spatial data from such a keyword search. The same method is used when searching for spatial data as with other keyword searches, with the key difference being to root the BFS in a spatial vertex.

This BFS method can be implemented with a temporal dimension instead of spatial. When changing the data type of the root from spatial to temporal, the traversal method stays the same, but the root used is matched with temporal input instead of spatial input. The result tree for a temporal search can be ranked using the same methods as used for spatial search.

Combining spatial and temporal data makes spatiotemporal search possible. These searches will be rooted in a vertex containing spatial and temporal data matching the input, as defined in 4.1. Because the root needs to contain both the spatial and the temporal input, the subset of vertices containing this wil be small. This makes spatiotemporal searches quick to execute, but the accuracy will suffer. A method that could be implemented to increase the hit rate of these searches would be to change the criteria for one of the two dimensions. The search could be rooted in either space or time, and then look for the other dimension as a special case while traversing.

Spatiotemporal searches rooted only in time could be accomplished by discarding all result trees that are unable to find a vertex inside the queried location. To discover vertices inside a location the “isLocatedIn” predicate would be followed. This predicate would be followed from all vertices that is connected to another vertex with that predicate. Using this predicate to move upwards in the hierarchy it creates while looking for the queried place will determine if the discovered vertex is located inside the search area. A maximum distance of three should be used when following this predicate, as this will traverse the hierarchy up to at least country level, and possibly up to the earth vertex. Since we are looking for a specific vertex, and only looking at the vertices above in the hierarchy this will not add more than three extra visited vertices for each vertex connected with the isLocatedIn predicate. This will not increase the time used by any significant margin compared to the regular temporal search.

Rooting spatiotemporal search in a spatial vertex and treating the temporal dimension as a special case does not require any extra traversal. During traversal, the search would look for vertices connected by a temporal predicate. This can be done by comparing predicates with a predetermined list of temporal predicates, checking the if the type of the object is “xsd:date” or “xsd:dateTime”, or it can be done by checking if the “rdfs:range” of the predicate is “xsd:date” or “xsd:dateTime”. Checking the predicate would add an extra traversal to each predicate connected to a vertex since the rdfs:range is treated as any other vertex. Using a predetermined list requires more knowledge of the data set, and adds some small overhead. Checking the type of the objects is the fastest since this would not require any extra traversal, and carries little overhead. This would not require any extra knowledge about the data set, assuming that no extra user added date types are created for the graph.

When traversing an rdf graph, the predicates chosen will have a great effect on the time used. Knowing the data searching through will help when choosing the predicate to follow. When choosing what predicates to use in the search, the goal should be to minimize the amount of unnecessary nodes found and followed. When removing predicates, the obvious downside is the possibility of missing some results, and decreasing the accuracy.

BFS search can take ua a lot of memory. This is because all nodes have to be stored while searching. It is possible to limit the maximum distance a node can be from the root node, which in turn will limit the amount of nodes visited, and reduce the memory needed. When testing without any form of pruning, the computer would run out of memory. Because of this, no results are gathered from a method following all edges of each node. Following all edges on each node would also lead to highly connected category nodes being discovered. With more than 120 million nodes, and 350.000 classes, each node would on average be connected to more than 340 other nodes, from the classes alone. Following this with a depth of 3, the average search would visit more than 40 million nodes.

From the results we can see that reducing the amount of nodes traversed results in significantly less time used for a query. By incorporating some natural language processing methods, and inferring structure from the users keywords, similar to those seen in [19, 11], the amount of nodes can be reduced. Using natural language processing to infer a category would make it possible to select a small set of predicates to follow. Such a category could also be used with the taxonomy of YAGO to only select nodes that fits.

Reduction of root nodes is the most important for increasing speed. To be able to reduce the roots while maintaining accuracy, using more sophisticated natural language processing is an option. Implementing a system for inferring structure from the keywords would make it possible to find connections where the object and predicate is related to the query. Using the taxonomy from YAGO, this is possible to implement.

Natural language processing also have the possible benefit of pruning some predicates. This would reduce the amount of vertices visited, and if this can be done without reducing accuracy, it would greatly increase the effectiveness when searching through RDF graphs.

Using an index for keywords can be used to find shortest paths without exploring the entire graph. Such a shortest path should be the best result for a given keyword, but depending on how it is implemented, it might not find the best solution for vertices containing multiple keywords. This is because a vertex containing more than one keyword at a higher depth may score better than multiple vertices containing one keyword each at a lower depth. The score would depend on the amount of keywords in a single vertex, and the difference in depth. If even a single vertex is found at the same depth as the multi-keyword vertex, the result tree with the fewest vertices will score better.

6.2 limitations

Some temporal vertices in YAGO use a so called "wildcard date". These dates contain a "#" symbol for parts of the date, making it impossible to directly query such dates when selecting roots. This means that the temporal search have some possible vertices missing from the set of root nodes used when traversing the graph. This could have been remedied by creating two new temporal vertices for such dates, one start date indicating the lowest possible value a wildcard date could have, and an end date indicating the highest possible value a wildcard could have.

When searching spatiotemporal data the wildcards have been used. This was done by adding some logic comparing the wildcards found connected to spatial vertices to the date range in the search. By doing this, the spatiotemporal search should be more accurate, and should retrieve some nodes that a regular temporal search would not retrieve.

7 Conclusion and Future Work

7.1 Contributions

This thesis had the goal to research methods for keyword searches on large spatiotemporal RDF graphs. This was broken down into three research questions.

Research question 1 *How can spatiotemporal data be integrated into exiting keyword query methods for RDF data.*

In the section 2 previous methods used for keyword searches on RDF graphs was investigated. Using methods from other research spatiotemporal keyword search have been implemented. The effectiveness of the search can be further improved, and implementing methods such as natural language processing for categorizing the keywords could further increase speed, without being detrimental to accuracy.

Research question 2 *What methods can be used to achieve greater speed and accuracy for searches on RDF data.*

This thesis looked at the parameters of a search that could increased speed, and decreased accuracy. When increasing speed, the main goal should be to reduce the amount of root vertices used as starting points for a search. Reducing the edges followed will also have a great effect on the speed. To be able to do this, processing the query, so that a structure can be created from the keywords is a possibility.

Research question 3 *How do spatial and temporal RDF query methods differ from from other query methods.*

Using a BFS for traversing the graph, a spatiotemporal search does not need to differ from spatial or temporal search. When selecting the roots used as starting points for the search, a spatiotemporal search needs to find roots that fall into both the spatial dimensions of the search, and the temporal. This makes the total set of roots used for a spatiotemporal search less than if the search was done just on one dimension.

7.2 Future Work

Combining structured queries natural language processing with graph traversal can yield great benefits for searching graphs.

Indexing and traversal form keywords and not just nodes is a possibility for increasing the effectiveness of searches. Adding toi this, finding shortest paths in the graph is another method for finding results.

Bibliography

[1] Dave Beckett and Brian McBride. Rdf/xml syntax specification (revised). *W3C recommendation*, 10(2.3), 2004.

[2] Dan Brickley, Ramanathan V Guha, and Andrew Layman. Resource description framework (rdf) schema specification. 1999.

[3] John Davies, Rudi Studer, and Paul Warren. *Semantic Web technologies: trends and research in ontology-based systems*. John Wiley & Sons, 2006.

[4] DBpedia. Dbpedia. URL `https://wiki.dbpedia.org`.

[5] Stefan Decker, Prasenjit Mitra, and Sergey Melnik. Framework for the semantic web: an rdf tutorial. *IEEE Internet Computing*, 4(6):68–73, 2000.

[6] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. 09 2016.

[7] Shady Elbassuoni and Roi Blanco. Keyword search over rdf graphs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 237–242. ACM, 2011.

[8] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, 9(1): 77–129, 2018.

[9] Claudio Gutierrez, Carlos A Hurtado, and Alejandro Vaisman. Introducing time into rdf. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):207–218, 2006.

[10] Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. Blinks: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316. ACM, 2007.

[11] Vanessa Lopez, Victoria Uren, Enrico Motta, and Michele Pasin. Aqualog: An ontology-driven question answering system for organizational semantic intranets. *Journal of Web Semantics*, 5(2):72 – 105, 2007.

[12] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*, January 2013.

[13] Brian McBride. Jena: A semantic web toolkit. *IEEE Internet computing*, 6(6):55–59, 2002.

[14] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon Bull Rev*, 21:1112–1130, 2014.

[15] Ferozuddin Riaz and Khidir M Ali. Applications of graph theory in computer science. In *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, pages 142–145. IEEE, 2011.

[16] Jieming Shi, Dingming Wu, and Nikos Mamoulis. Top-k relevant semantic place retrieval on spatial rdf data. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 1977–1990. ACM, 2016.

[17] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *16th International Conference on the World Wide Web*, pages 697–706, 2007.

[18] Jonas Tappolet and Abraham Bernstein. Applied temporal rdf: Efficient temporal querying of rdf data with sparql. In *European Semantic Web Conference*, pages 308–322, 2009.

[19] T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In *2009 IEEE 25th International Conference on Data Engineering*, pages 405–416, March 2009.

[20] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference*, pages 1–6, 2010.

[21] Denice Worthington and Paul Nation. Using texts to sequence the introduction of new vocabulary in an eap course, 1996.

Appendices