Hans Sandbu

Spatio-temporal 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 keyword searching in spatiotemporal graph data, with the goal of finding what aspects are needed for accuracy and speed. The goal is to recreate previous methods used, and through testing determine what aspects of an algorithm have most impact on the speed and accuracy of finding and ranking results.

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, 5th May 2020

Contents

List of Figures

List of Tables

1 Introduction

Creating efficient structures for computers to interpret information opens up many new areas for information retrieval. One method is structuring the information as a graph with vertices connected by edges (G(v, e)). In such a structure each vertex (v) can hold a small piece of information, and the edge (e) represents a relation. When information is modeled in this way, it can be called a knowledge graph (KG).

Graph based models allows complex relations to be modeled. Finding relations in a graph is also computationally cheaper than other models. Using standardized models such as the Resource Description Framework (RDF), which is created specifically for graph based data, has made it possible to develop efficient storage and retrieval solutions for graph data. One such solution is the framework Jena (Wilkinson et al., 2004) for Java.

Using RDF projects like YAGO (Yet Another Great Ontology) (Suchanek et al., 2007) have created graphs containing spatial and temporal data (spatiotemporal). This makes it possible to apply reason to the data set, and derive new information from what is already there. This is a feature that typically separates a KG from an ontology (Färber et al., 2017). In a graph, spatiotemporal data will be described using at least 2 vertices, one for a start date, and one for a place. When modeling time intervals, multiple vertices are needed.

There are few applications using RDF as an information storage system. Some of the more well known are DBPedia, YAGO, 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. The applications built with RDF usually don't have much utility. A common utility built on top of RDF data sets are tools used to describe social relations, though there have been developed applications such as MusicBrainz that use RDF for relations between songs, artists, albums and other tags given to entities.

Creating methods that allows for easy access to information in RDF data makes it possible to create applications with more utility. Keyword and text search in RDF data is still something being developed and improved. Using frameworks such as Jena, it is possible to build solutions for keyword search, and spatiotemporal search. In this thesis (paper?) methods for efficient keyword search on spatial and temporal data will be explored.

1.1 Background and Motivation

Traditionally relationships between pieces of data have been difficult for computers to model. Regular SQL databases requires expensive joining to be able to display relationships between entities. 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 *With this thesis the goal is to determine if and how spatiotemporal keyword searching in large scale RDF graphs is possible to do in fast and accurate.*

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.*

Research question 2 *What methods can be used to create more effective queries on RDF data.*

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

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.

1.4 Thesis Structure

2 Background Theory

2.1 Definitions

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.

Root: *Any GeoEntity node used as a start when traversing the graph, and as the place node in a result tree*

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* $\frac{\sum \mathcal{F}_h}{n*(nd+1)}$

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.

2.1.1 Knowledge Base

The term knowledge base can have many definitions, often because the terms knowledge base, ontology, and knowledge graph is used interchangeably (Ehrlinger and Wöß, 2016). In this paper the term knowledge graph will be defined as follows: "A knowledge base is a data set with some formal semantics. This could include multiple axioms, definitions, rules, facts, statements, and other primitives." (Davies et al., 2006)

2.1.2 Ontology

Like knowledge base, ontology does not have one clear definition. This is because the term have many different interpretations, and is often confused with other terms (Feilmayr and Wöß, 2016). Most definitions of ontologies say that ontologies represent a schema, basic theory, or conceptualization of a domain, which knowledge bases do not. (Davies et al., 2006) To differentiate ontologies and knowledge bases in this paper the definition of an ontology will be "An ontology is an extended knowledge base that allows for semantic modeling of knowledge." With this definition, ontologies can be considered a specialized form of knowledge bases.

2.1.3 Knowledge Graphs

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.” (Ehrlinger and Wöß, 2016) This definition is encompasses multiple technologies. In this paper we will use the more specific definition that fits the data used “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 ∈ U ∪ B, a predicate p ∈ U, and an object U ∪ B L. An RDF term is either a URI u ∈ U, a blank node b ∈ B, or a literal l ∈ L” (Färber et al., 2017)

Subject

In a 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.

Predicate

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.

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.

2.1.4 Facts and entities

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.

2.2 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.

2.2.1 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.

2.2.2 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(Suchanek et al., 2007) and an ontology (Mahdisoltani et al., 2013), but is often described as a knowledge graph by others. In this paper Yago is described as a knowledge graph.

Temporal data

Yago have many entities with facts describing date and spatial data.(Suchanek et al., 2007) 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.(Suchanek et al., 2007) 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.(Suchanek et al., 2007)

Spatial data

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".

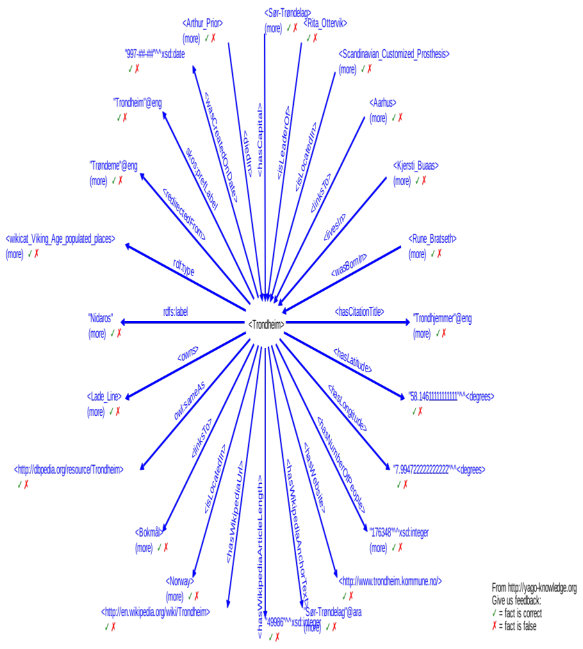

<Trondheim>
<Sør-Trøndelag> (more) ✓ ✗
<Arthur_Prior> (more) ✓ ✗
<Rita_Ottervik> ✓ ✗
"997-##-##"^^xsd:date ✓ ✗
<Scandinavian_Customized_Prosthesis> (more) ✓ ✗
"Trondheim"@eng ✓ ✗
<Aarhus> (more) ✓ ✗
"Trønderne"@eng (more) ✓ ✗
<Kjersti_Buaas> (more) ✓ ✗
<wikicat_Viking_Age_populated_places> (more) ✓ ✗
<Rune_Bratseth> (more) ✓ ✗
"Nidaros" (more) ✓ ✗
"Trondhjemmer"@eng (more) ✓ ✗
<Lade_Line> (more) ✓ ✗
"58.1461111111111"^^<degrees> ✓ ✗
<http://dbpedia.org/resource/Trondheim> ✓ ✗
"7.99472222222222"^^<degrees> ✓ ✗
<Bokmål> (more) ✓ ✗
"176348"^^xsd:integer (more) ✓ ✗
<Norway> (more) ✓ ✗
<http://www.trondheim.kommune.no/> ✓ ✗
<http://en.wikipedia.org/wiki/Trondheim> ✓ ✗
"49986"^^xsd:integer ✓ ✗
"Sør-Trøndelag"@ara (more) ✓ ✗
<wasCreatedOnDate>
<diedIn>
<hasCapital>
<isLeaderOf>
<isLocatedIn>
<linksTo>
skos:prefLabel
<redirectedFrom>
<livesIn>
rdf:type
<wasBornIn>
rdfs:label
<hasCitationTitle>
<owns>
<hasLatitude>
owl:sameAs
<hasLongitude>
<linksTo>
<hasNumberOfPeople>
<isLocatedIn>
<hasWebsite>
<hasWikipediaUrl>
<hasWikipediaAnchorText>
<hasWikipediaArticleLength>
From http://yago-knowledge.org
Give us feedback:
✓ = fact is correct
✗ = fact is false

Figure 2.1: Trondheim as a subject in YAGO

2.2.3 DBPedia

2.2.4 Jena

Apache Jena is a set of tools for working with RDF and other semantic web, and linked data. The framework contains tools for querying with SPARQL, 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.

2.3 Introducing Figures

2.4 Introducing Tables in the Report

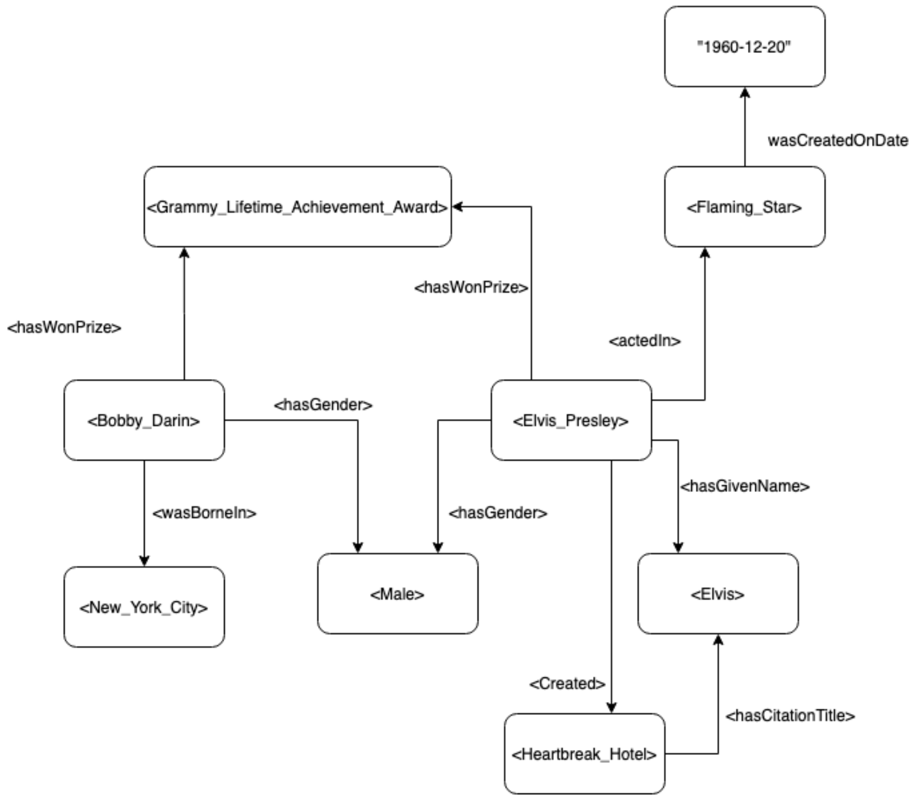


Figure 2.2: Connections in Yago around Elvis

3 Related Work

Keyword search on RDF graphs often follows a set of common strategies. 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 (He et al., 2007) 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 (Elbassuoni and Blanco, 2011) 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 (Lopez et al., 2007). 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 (Tran et al., 2009). 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 (He et al., 2007; Elbassuoni and Blanco, 2011; Tran et al., 2009) 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 Elbassuoni and Blanco (2011) 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, Tran et al. (2009), 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.

keyword searching and ranking spatial RDF graphs is quite similar to regular RDF keyword searching. Shi et al. (2016) outlines methods that incorporates spatial data into the search. These methods is similar to some of those previously mentioned here (Tran et al., 2009; Elbassuoni and Blanco, 2011). 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.

R-trees are designed around the spatial dimension, but some work on modifications to include a temporal dimension has been done. Some research done on spatiotemporal tree structures include (Tao et al., 2003; Šaltenis et al., 2000). Most of the spatiotemporal trees are however created to be able to query and index moving or frequently updated objects. To include a temporal dimension some differences in the tree structures are made. Most approaches builds on the R-tree, but modifies the tree in different aspects. For the purpose of this thesis a different tree structure is not needed.

4 Architecture

4.1 Search

4.1.1 BFS search

The most basic method of finding a match for keywords is using a breadth first search (BFS) (He et al., 2007; Shi et al., 2016). 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. There is however no guarantee that this set of vertices contains any spatial or temporal data. If this is the case, the search will continue until the a vertex containing spatial and temporal data is found. The first set of vertices containing all the keywords, a place, and time will be the best result using this search. The BFS search can however contain multiple times, or multiple places. To determine what time or what place best fits the query, a separate ranking algorithm can be used. BFS is also a time consuming algorithm and is used as a proof of concept and baseline in this thesis.

Algorithm 1: GetFullResultTree(p, t, Qt)

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

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 4.2 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 documnet 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.

Algorithm 2: minimum spanning graph

```
 1: procedure MINIMUM For each node in hits:
 2: for each minNode in min list:
 3: if node has same hits and higher depth than minNode:
 4: Break
 5: if node and minNode is equal in depth and hits:
 6: continue
 7: if node contains other words than minNode:
 8: add node
 9: remove potential common words from minNode
10:
```

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.

- Find the minimum spanning tree for each of the root nodes containing the maximum amount of the query words.

- Rank based on 1. Query words hit 2. Nodes in the tree 3. depth of the tree
The final piece is to rank the minimum subgraphs. This is done using the nr. 2 formula found in Shi et al. (2016).

4.2 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.

4.2.1 Predicate pruning

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.

4.2.2 Unqualified place pruning

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.

4.2.3 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 search methods. when comparing search methods, speed and accuracy will be the metrics used. Three methods for searching is used, one following all gathering all objects connected to an input subject, regardless of the predicate. One method following Yago predicates leading directly to a Yago URI. And the final method will follow the same predicates as the second, and will also add a condition where all nodes must match at least one query word.

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 query 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.

BFS

BFS with pruning

5.2 Experimental Setup

5.3 Data set

When selecting a data set an important feature was accurate spatial and temporal data. For this Yago was selected. Yago allows for download of subsets of data, and contains all the necessary data. The high accuracy in Yago was also a benefit of this data set. From Yago, the data selected was the entire Yago taxonomy, the entire Yago core, and from Genomes the sets GeonamesOnlyData, GeonamesClassIds, GeonamesGlosses, GeonamesTypes, and GeonamesClasses were selected. These sets makes it possible to build a complete graph of the entities in Yago, but keeps the disk usage as low as possible.

- Short description of each of the parts
- preprocessing: replace non-unicode chars, replacing space with underscore in URIs, terminate triplet where missing termination, remove double quotes in URIs, remove backslash unless legal escape sequence
- tdbloader for persistent queryable storage
- Structure of graph? Here or in Yago description?

5.4 Queries

When selecting the keywords used in queries, a semi random selection method was used at first. This method created three pools of words, rare words (less than 1000 occurrences) uncommon words (less than 100 000 occurrences) and common words (more than 100 000 occurrences). The three pools where created by stemming all words, then counting occurrences. From each pool of words, a set of 100 where randomly selected, and from those 100, 10 where selected manually to ensure a range of occurrences, and to ensure no stop words, misspelled words or other errors were in the final set of keywords.

A second set of words were also chosen to ensure a larger hit rate. This set was a random selection of 20 words from the top 150 words. From the 20 random words 10 where manually selected, the 150 word number was selected based on Zipfs law Piantadosi (2014).

When selecting places for spatial queries, a set of 20 random places were selected from the YagoGeonamesOnlyData set, and from that set, 5 were chosen manually for the final set. In the final set, three places where added manually, Oslo, London, and New York City were added to ensure variation in placement and node connections.

Generating random queries can create results that will not have any hits.
Find a good method of creating queries.

5.5 Pruning

A simple method to improve performance when searching in large data sets is to remove data that is unnecessary to search.

5.5.1 Predicate pruning

Predicate pruning is the process of eliminating predicates from the search. The taxonomy of Yago creates assigns predicates to one or more group or class. By selecting specific groups or classes it is possible to follow the types of data searched for, while eliminating a lot of other data. When pruning based on classes or groups the type of data returned must be known beforehand, to ensure that no possible hits is overlooked.

Another possible method for pruning is to specify the exact predicates to follow. This will remove a lot of data, but is possibility when searching for something specific. This will also greatly increase the speed of the search, and reduce the memory usage for the search.

5.6 Experimental Results

5.7 BFS search

BFS search without any pruning and search depth of 3 were unable to finish due to memory constraints and taking too long. Search with a depth of 2 may finish, depending on on the connectedness of the nodes discovered, and the amount of root nodes found in the search.

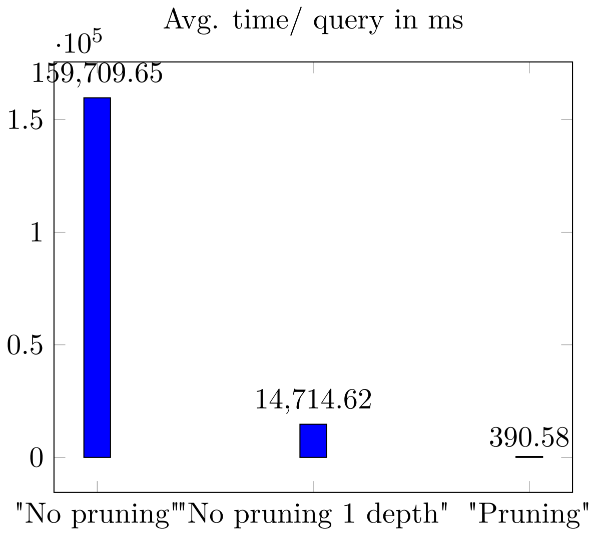
5.7.1 Time

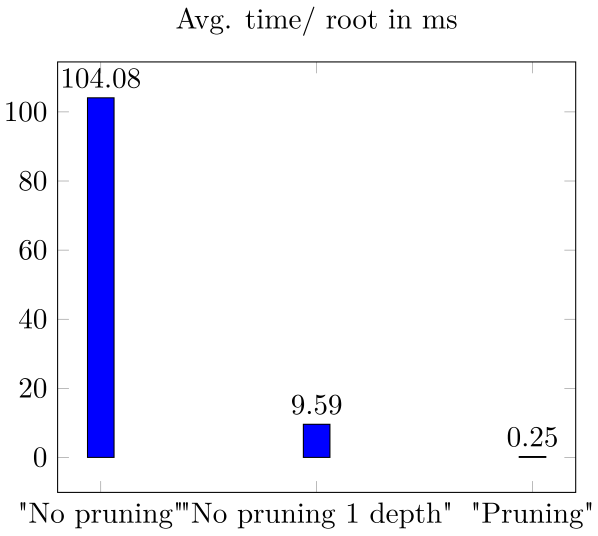
When running the query, two sets are displayed, one with mostly random input data, and one with more closely selected data.

Because most of the randomly selected places had few possible root nodes, the table contains only the manually selected.

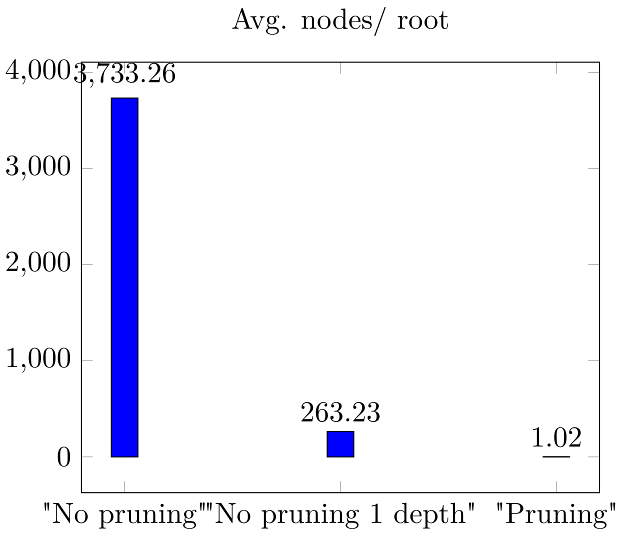
- time/query
- time/root
- roots found/ query

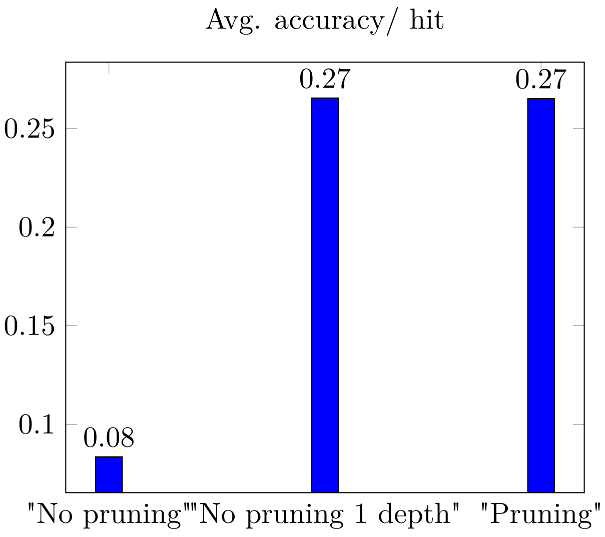
- accuracy/ result
- avg highest accuracy/ query

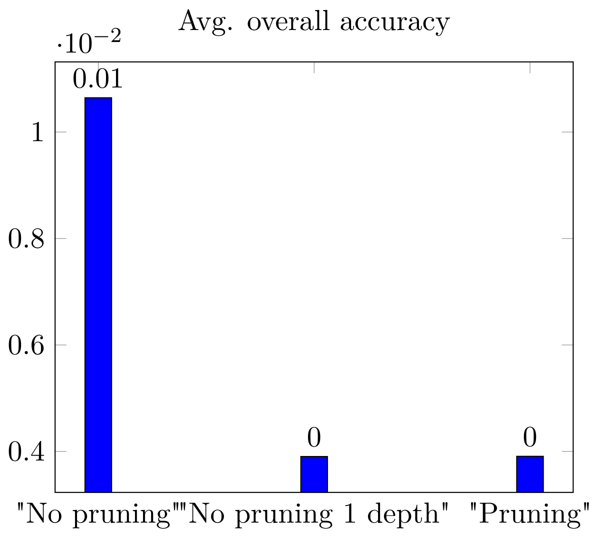


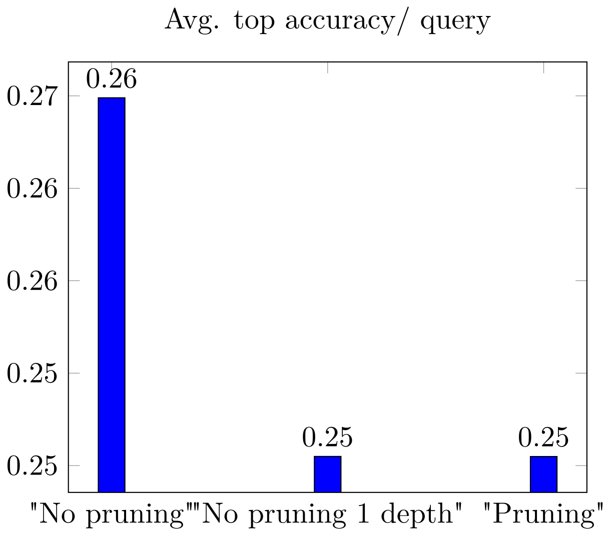


5.7.2 Accuracy






Avg. overall accuracy
$\cdot 10^{-2}$
0.01
0
0
1
0.8
0.6
0.4
"No pruning"
"No pruning 1 depth"
"Pruning"


Avg. top accuracy/ query
0.26
0.25
0.25
0.27
0.26
0.26
0.25
0.25
"No pruning"
"No pruning 1 depth"
"Pruning"

6 Evaluation and Discussion

6.1 Evaluation

6.2 Discussion

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.

- limitations
- Effectiveness of pruning
- choice of query terms
- Root nodes

7 Conclusion and Future Work

7.1 Contributions

7.2 Future Work

Bibliography

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

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

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, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063615. URL `http://doi.acm.org/10.1145/2063576.2063615`.

Christina Feilmayr and Wolfram Wöß. An analysis of ontologies and their success factors for application to business. *Data & Knowledge Engineering*, 101:1 – 23, 2016. ISSN 0169-023X. doi: https://doi.org/10.1016/j.datak.2015.11.003. URL `http://www.sciencedirect.com/science/article/pii/S0169023X1500110X`.

Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, 9:1–53, 03 2017. doi: 10.3233/SW-170275.

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, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-686-8. doi: 10.1145/1247480.1247516. URL `http://doi.acm.org/10.1145/1247480.1247516`.

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. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2007.03.003. URL `http://www.sciencedirect.com/science/article/pii/S1570826807000145`. Software Engineering and the Semantic Web.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, United States, January 2013. URL `https://hal-imt.archives-ouvertes.fr/hal-01699874`.

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

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, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3531-7. doi: 10.1145/2882903.2882941.

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.

Yufei Tao, Dimitris Papadias, and Jimeng Sun. The tpr*-tree: An optimized spatio-temporal access method for predictive queries. In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, VLDB '03, pages 790–801. VLDB Endowment, 2003. ISBN 0-12-722442-4. URL `http://dl.acm.org/citation.cfm?id=1315451.1315519`.

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. doi: 10.1109/ICDE.2009.119.

Kevin Wilkinson, Craig Sayers, Harumi Kuno, Dave Reynolds, and Jena Database. Efficient rdf storage and retrieval in jena2. *Proceedings of SWDB*, 02 2004.

Simonas Šaltenis, Christian Jensen, Christian (codirector, Michael Böhlen, Curtis Dyreson, Heidi Gregersen, Dieter Simonas, Janne Skyt, Giedrius Slivinskas, Kristian Torp, Richard (codirector, Bongki Moon, Michael Soo, Amazon Com, and Andreas Timeconsult. R-tree based indexing of general spatio-temporal data. 01 2000.

Appendices