Semantic data mining and linked data for a recommender system in the AEC industry

Petrova, Ekaterina Aleksandrova; Pauwels, Pieter; Svidt, Kjeld; Jensen, Rasmus Lund

Published in:
Proceeding of the 2019 European Conference on Computing in Construction

DOI (link to publication from Publisher):
10.35490/EC3.2019.192

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

2019 EC³
2019 European Conference on Computing in Construction
Chania, Crete, Greece
July 10-12, 2019

SEMANTIC DATA MINING AND LINKED DATA FOR A RECOMMENDER SYSTEM IN THE AEC INDUSTRY

Ekaterina Petrova[1,*], Pieter Pauwels[2], Kjeld Svidt[1] and Rasmus Lund Jensen[1]
[1]Department of Civil Engineering, Aalborg University, Aalborg, Denmark
[2]Department of Architecture and Urban Planning, Ghent University, Ghent, Belgium

Abstract

Even though it can provide design teams with valuable performance insights and enhance decision-making, monitored building data is rarely reused in an effective feedback loop from operation to design. Data mining allows users to obtain such insights from the large datasets generated throughout the building life cycle. Furthermore, semantic web technologies allow to formally represent the built environment and retrieve knowledge in response to domain-specific requirements. Both approaches have independently established themselves as powerful aids in decision-making. Combining them can enrich data mining processes with domain knowledge and facilitate knowledge discovery, representation and reuse. In this article, we look into the available data mining techniques and investigate to what extent they can be fused with semantic web technologies to provide recommendations to the end user in performance-oriented design. We demonstrate an initial implementation of a linked data-based system for generation of recommendations.

Introduction

Building data: BIM and semantic web technologies in a sensor world

Recent years have presented significant research efforts accentuating the environmental impact of the built environment and methods for its mitigation. That has amended design practice and has made it strive towards implementing sustainability principles as fundamental and not merely complementary. Simultaneously, the rapid technological developments have allowed powerful computational techniques to emerge in support of architectural design and engineering. Such technologies allow to represent buildings semantically (Pauwels et al. 2017) and discover implicit knowledge about their performance through pattern recognition and knowledge discovery techniques (Fayyad et al. 1996). With regards to data representation in Architecture, Engineering and Construction (AEC), Building Information Modelling (BIM) allows the creation of semantically rich building models (Sacks et al. 2018).

Recently, semantic web technologies (Berners-Lee et al. 2001) have received major attention in the attempt to break open the isolated information silos and connect the semantically rich building data with other meaningful data about the building, its occupants, environment, etc. These further reaching semantic models are the building blocks of Linked Building Data (LBD) and provide a decentralized source of information (Pauwels 2014). On the other hand, Building Monitoring Systems (BMS) play an essential role in building operation, by allowing the collection of operational data through a myriad of sensors and devices (Fan et al. 2015). Advanced analytical methods are hereby of high value, as they help uncover hidden knowledge in the data, and highlight its potential to the future of building design and performance improvement (Fan et al. 2015, Miller et al. 2018).

Despite the availability of knowledge bases, many of the decisions taken during the design process are based on 'rules of thumb' and previous experience (Heylighen et al. 2007), and not on data and evidence contained in building performance, BIM models or LBD knowledge graphs. If such data were used more efficiently, significant potential would be uncovered in reaching performance targets currently associated with gaps between design and actual performance (de Wilde 2014). The target of this research effort is to bring knowledge from previous projects into future design environments to achieve both a sustainable end product and a holistic sustainable design process. Previous works also investigated how Knowledge Discovery in Databases (KDD) (Fayyad et al. 1996) can be used to retrieve patterns and association rules from available building data (Petrova et al. 2018*a*,*b*). These works also showed how it is possible to build a knowledge graph that includes (1) semantically rich building data (topology, properties, etc.), (2) 2D and/or 3D geometry, (3) sensor data, and (4) motifs and association rules obtained from the sensor data. The resulting graph provides a valuable resource for evidence-based design recommendations. Therefore, the objective of the current article is to investigate the potential of linked data-based recommendation retrieval in the design environment, including performance patterns discovered in sensor data, thereby utilizing the available

DOI:10.35490/EC3.2019.192

and ever-growing knowledge bases to achieve an evidence-based design process.

Linked data-based recommender systems for improving sustainable design decision-making

This work explores the possibility of building a system that relies on knowledge disocvered in building data and stored in knowledge graphs to make recommendations towards the design team. Considered here is evidence-based feedback in response to design requirements, yet the recommender system is conceived as user-centered and can provide any feedback requested by querying the available knowledge base(s). Recommender systems can be subdivided in content-based and graph-based (Musto et al. 2017), where content-based systems provide recommendations based on direct similarity and graph-based ones link user nodes to user-tailored recommendations.

Several research efforts investigate recommender systems based on linked data and the wealth of data provided by the Linked Open Data (LOD) cloud[1](Oliveira et al. 2017, Musto et al. 2017). Research in the area of LOD-based recommendations takes its roots in the field of ontology-based recommender systems introduced by Middleton et al. (2004). When linked data and ontologies are used for the disambiguation of content, recommendation systems become semantics-aware (Boratto et al. 2017). The use of linked data for user-centered recommendations was introduced by Passant (2010), who proposed a recommender system based on semantic similarity calculations. This research relies on a set of measures to compute the semantic distance in linked data, thus exploiting the abundance of links among the resources. Recent works (Oliveira et al. 2017, Boratto et al. 2017) typically follow the software architecture displayed in Figure 1 where user profiling is on focus.

Recommender systems nowadays are usually associated with user profiling and recommendations based on previous interactions, social relations, likes, etc. The aim is to match the user's demands (profile) with the highest possible level of similarity, while still diversifying the recommendations. In the case of building design, the similarity matching aspect should also be the starting point, but it should be equally balanced with diversification driven by design and performance requirements. For example, if a user indicates high interest in residential nearly zero-energy buildings (NZEB), the recommender system should also be able to suggest other NZEB building types or other residential building types, etc. Of course, the richer the original dataset, the easier it is to obtain and make alternative recommendations.

Recommendation engines are not unknown to the AEC industry. However, these usually suggest predefined objects hosted in a database when a certain level of similarity with the current design is achieved (content-based). As a result, one of the fundamental goals of this research is to investigate the level of feasibility for application of linked data-based recommendations utilizing dynamic knowledge bases in changing context. The dynamic knowledge bases can be new buildings projects, which may include continuous incoming streams of sensor data and new LBD graphs. The changing context then refers to continuously updating user profiles.

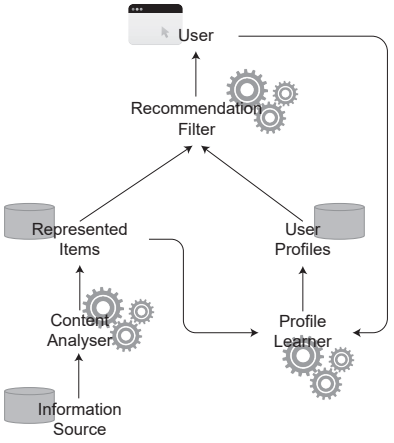

Figure 1: Semantics-aware content-based recommender system (based on Boratto et al. (2017))

To achieve the above-stated objectives, this paper starts with a state of the art review in the areas of KDD, semantic web technologies and data stream processing. The article continues with the approach uses to achieve the objectives of the current study. We then outline the necessary steps towards a linked data-based recommender system for improvement of decision-making in sustainable building design and perform initial tests. Finally, the paper discusses the results, presents the main conclusions and outlines future work.

State of the Art

Knowledge Discovery in Databases (KDD) according to data type and purpose

Fayyad et al. (1996) define KDD as an overall process, in which knowledge is the end product of data-driven discovery. They outlinee five main steps, namely selection, pre-processing, transformation, data mining and interpretation/ evaluation of the results. In that context, Hand et al. (2011) define data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarise the data in novel ways so that data owners can*

[1]https://lod-cloud.net/

fully understand and make use of the data". Fayyad et al. (1996) also summarise six main data mining categories, i.e., classification, clustering, association rule mining, regression, summarization and anomaly detection. Han et al. (2012) divide these into two main categories: predictive and descriptive. Descriptive analystics use data aggregation and mining to provide insight into the past and make it interpretable by humans. Predictive analytics use statistical models and forecasting to understand the future and provide actionable insights. With regards to the input, Lausch et al. (2015) distinguishes between (numerical and categorical) data, text, web, media, time series and spatial data mining.

Knowledge discovery in Architecture, Engineering and Construction

Petrova et al. (2018*c*) provide an extensive definition of KDD approaches according to the type of building data (semantic BIM data, geometric data, sensor data, etc.) and the knowledge discovery purpose. Due to the abundance of spatio-temporal data, the AEC industry can benefit from mining temporal data (time series) and spatial data. Shekhar et al. (2010) rightfully indicates that extracting interesting patterns and associations from such complex and multidimensional data with plenty of dependencies and spatio-temporal correlations is more difficult than mining traditional numeric and categorical data. In AEC, spatio-temporal data mining approaches can be valuable in cases where spatial data is augmented with time series data from sensor networks in buildings.

Data mining applications for building performance improvement and sustainable building design usually relate to energy use and demand prediction (Wang & Srinivasan 2017), prediction of occupant behavior (D'Oca & Hong 2014), fault detection for building systems (Cheng et al. 2016), improvement of building operation and control strategies (Xiao & Fan 2014), as well as discovering and explaining energy use patterns (Miller et al. 2015). Other researchers have investigated the use of semantic data modelling, neural networks and data mining for building energy management (McGlinn et al. 2017). As can be seen from these categories, the use of KDD is usually related to the improvement of building operation. Using such approaches to improve future building design processes have not been investigated in such detail.

Limitations in the application of data mining

"Classic" data mining techniques typically focus on isolated "silo" data. As stated by Lausch et al. (2015), in such cases, the conclusions remain limited and do not span interdisciplinary and complex data. Additionally, data selection and treatment resides in the hands of the analyst, who is responsible for variable selection and data preparation to fit the needs of the mining algorithms. In case of incorrect decisions, the results can be influenced negatively, e.g. hidden patterns and novel knowledge may not be discovered or registered.

Therefore, Lausch et al. (2015) propose to mine data using linked data technologies. Such an approach allows opening silos and integrating data across disciplines, and provides an opportunity for analysis of interdisciplinary datasets. This broad overview can lead to insightful analyses, especially in a semantically rich domain such as AEC. Nevertheless, how these analyses are obtained is very different from the methods used in data mining, in the sense that the linked data realm is governed by queries and rules. These methods can be considered graph mining or matching techniques, and therefore potentially similar to pattern recognition. However, the types of graphs and patterns used in semantic queries and rules are very different from the patterns uncovered using data mining techniques, and both should not be perceived as identical.

Knowledge Graphs, Linked Data and the Semantic Web

Further to the evolutions in KDD, a lot of progress has been made in the formalization of knowledge using web technologies. From a web of documents, the World Wide Web has evolved into a 'Web of Data' (LOD cloud) (Bizer et al. 2009). The term linked data was coined by Tim Berners-Lee in 2006[2] and has enabled worldwide publication of 5-star open data[3]. This implies defining data according to the Resource Description Framework (RDF)[4] data model and interlinking it with other RDF datasets available on the web. The Web of Data relies on ontologies so that data is typed and can easily be used in combination with query and rule languages such as SPARQL. Ontologies can be defined using RDFS and OWL[5] and give 'meaning' or 'semantics' to the data, thereby constituting the Semantic Web as conceived in Berners-Lee et al. (2001).

Due to their potential, linked data and semantic web technologies have received major attention in the AEC industry. A comprehensive overview of this topic can be found in Pauwels et al. (2017). Among the most notable initiatives is the early work on transforming the Industry Foundation Classes (IFC) into an OWL ontology (ifcOWL) (Pauwels & Terkaj 2016). The ifcOWL ontology was built to match the original EXPRESS schema as closely as possible, thus allowing a round-trip conversion process (lossless conversion). However, this has lead to a very big ontology, which resembles the IFC schema almost entirely, i.e., difficult to extend, complex, and not modular.

[2]http://www.w3.org/DesignIssues/LinkedData.html
[3]http://5stardata.info/
[4]http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
[5]http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

This led to research initiatives aiming at ontologies for Linked Building Data, which do not rewind to IFC, yet cover similar ground.

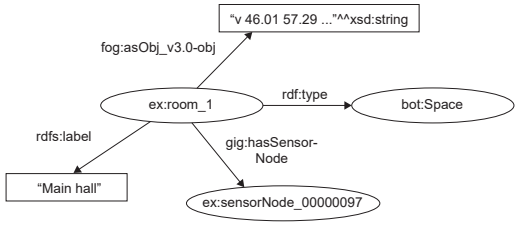

Figure 2: An example LBD graph.

An ecosystem of smaller domain ontologies is currently available, each covering parts of what can also be handled with IFC (Figure 2 and 3). A central Building Topology Ontology (BOT) Rasmussen et al. (2017) captures terms as 'Building', 'Site', 'Space', etc. and aims for standardisation of these terms within the W3C LBD CG. Starting from BOT, alignments with various domain ontologies (Schneider 2017) can then be made. As a result, the industry can rely on a modular set of ontologies, yet still have a stable standard at the core. Besides topology, other ontologies in the W3C LBD CG cover products, properties, and geometry McGlinn et al. (2019).

Semantic Data Mining

Standard data mining algorithms usually use statistical models on data to discover patterns and provide actionable insights. According to Lavrač et al. (2011), in these cases, data is treated as meaningless numbers and attribute values. In other words, data by itself does not convey any semantics and needs to be interpreted to present meaningful information, which is usually done by domain experts. Such processes are associated with an abundance of raw data, but the underlying knowledge is scarce. Considering that KDD and data mining are knowledge-intensive processes, they can significantly benefit from enrichment by domain knowledge and the relations between objects. As further stated by Lavrač et al. (2011), that can be achieved by adding semantic annotations to the data and use of domain ontologies. This concept has caused a paradigm shift in data mining, expressed in a transition from mining the raw data to mining the knowledge directly. An overview of how semantic web technologies can be used in data mining and KDD is given in Ristoski & Paulheim (2016).

The increased interest in the fusion of data mining and semantics has also highlighted the main technical challenges and opportunities that this union presents. For instance, classic data mining is powerful for extracting patterns and association rules from large traditional datasets. Yet, as Nebot & Berlanga (2012) state, the different nature of semantic data presents challenges, which cannot be tackled by traditional machine learning approaches, as they target mostly homogeneous data composed by transactions (sets of items). Since annotated data does not follow a rigid structure, instances, which are a part of the same class may still have a different structure. That causes a challenge of structural heterogeneity. Together with the heterogeneity of data sources, this leads to the necessity of dedicated approaches for pattern discovery in semantic data. This includes reasoning capabilities that allow inferring the implicit knowledge residing in the ontology itself (`subClassOf` relations, rules, inverse relations, etc.). For those reasons, researchers have engaged in defining the pathway towards effective association rule mining in knowledge bases (Barati et al. 2016).

Storing and processing sensor data

An important body of work in the semantic web domain, which is also of particular relevance in this paper, lies in the context of sensors and actuators. Sensor nodes are placed in precisely determined locations with a particular purpose of observation, thereby monitoring building use and performance in a real-time manner. This typically results in significantly large amounts of data, often captured in data lakes. Such data can be used in RDF graphs (Semantic Sensor Networks), and thus be directly included as separate modules complementing the modular LBD cloud. Example ontologies that can be used for this purpose are SOSA[6], SSN[7] and SAREF[8].

Calbimonte et al. (2012) state that the heterogeneity of sensor data sources and environments is an important issue related to the realization of a connected sensor world. Monitored data is usually represented in different ways by different networks, and data models and schemas differ as well. That leads to several compatibility and representation issues. To tackle those, research efforts propose various solutions such as semantic annotation of sensor data (Sheth et al. 2008), providing ontology-based access to data (Calbimonte et al. 2010), etc.

Storing the vast amount of data directly in the RDF graph typically leads to a "swollen" graph, and takes down query and reasoning performance. Hence, Petrova et al. (2018*a*,*b*) propose to maintain sensor data within their common non-RDF based data stores, yet link directly from the RDF graph to the web API providing access to the sensor data. When relying on web technologies for application development, these HTTP links can be consumed to give a custom and on-demand access to the raw sensor data. However, several studies suggest that further opportunities may arise from using SPARQL queries with streaming ex-

[6]http://www.w3.org/ns/sosa/
[7]http://www.w3.org/ns/ssn/
[8]http://ontology.tno.nl/saref/

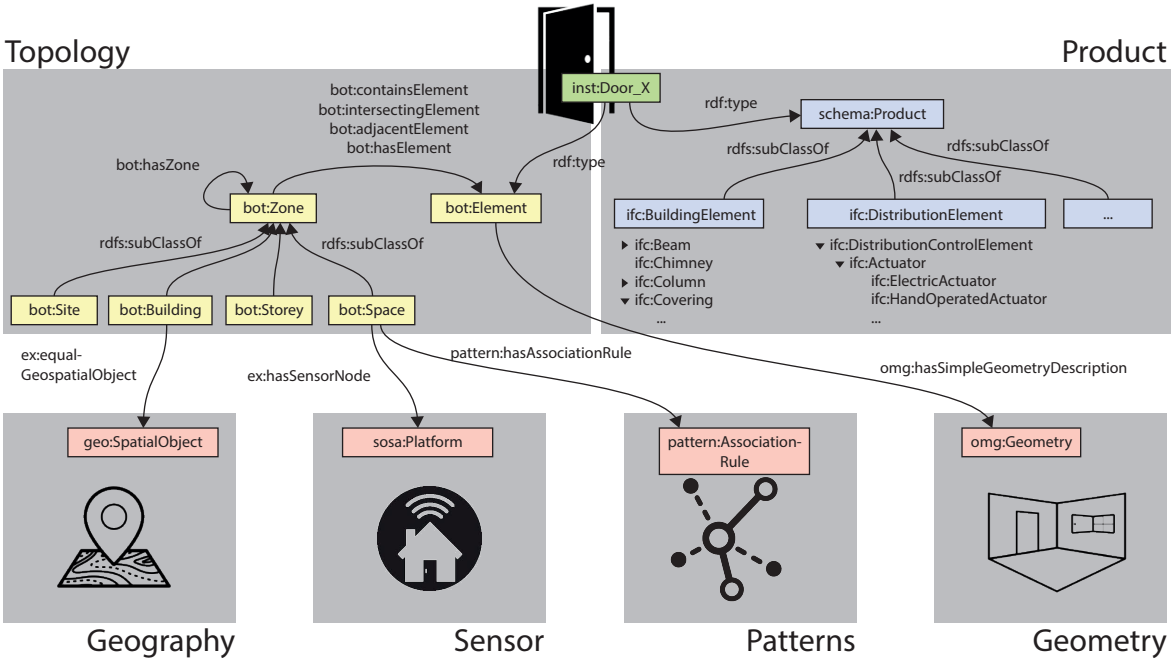

Figure 3: Conceptual overview of the modules and ontologies in a linked building data cloud, based on the work in the W3C LBD CG.

tensions to access observations (Calbimonte et al. 2012). RDF stream processing may give an opportunity to publish and analyze real-time data streams while avoiding the "swollen" graph issue and still make sensor data a part of the LBD knowledge graph. Della Valle et al. (2009) state that achieving that would require moving from storing data and querying it on demand ("one-time semantics") to using continuous queries ("continuous semantics"). Barbieri et al. (2010) state that focus needs to be put on "stream reasoning", i.e., making sense of multiple real-time heterogeneous data streams. Llanes et al. (2016) define three main stages in the publication of RDF streams, i.e. conversion from sensor data streams to RDF streams, storing RDF streams, and linking them with other data sources. That requires the selection of relevant ontologies, defining the mapping language for conversion, selection of continuous query languages (e.g. Continuous SPARQL (C-SPARQL) and SPARQLstream (Barbieri et al. 2010), (Calbimonte et al. 2012)) and choosing other appropriate datasets to link to.

Semantic Data Mining and Linked Data for a Recommender System in the AEC Industry

Conceptual framework

Based on the review, we conclude that in the implementation of the recommender system (1) knowledge graphs can be accessed using semantic queries, (2) sensor data can be mined with traditional data mining techniques, (3) semantic data mining can be performed on the LBD graph, and (4) RDF graph mining can also be used for pattern matching in combination with RDF stream processing.

Furthermore, a recommender system can rely on data sources both without and with explicitly embedded semantics. In the latter case, recommender systems rely directly on semantic analysis techniques, thereby directly exploiting the semantics in the linked data graph. In the current context, in which the modular LBD graphs consist of both graph data (topology and product data) and non-graph data (geometry and sensor data), both traditional and semantic data mining can be used. Mining of raw sensor data implies discovery of performance patterns by the use of classic data mining methods. The knowledge interpretation is strictly related to obtaining understanding about the performance through the discovered patterns, not through the raw data. The RDF frequent pattern discovery, on the other hand, is data structure oriented and considers the graph predicates instead of data values.

Applying these techniques results in the conceptual system architecture in Figure 4. The following sections explain this architecture focusing on (1) how patterns are discovered and added to the graph, (2) how user profiles can be built and benefit from the system, including feedback, and (3) how recommendations can be generated. We present an example for RDF pattern discovery in a semantic data stream by implementing a method suggested

by Belghaouti et al. (2016) and discuss its potential feasibility. Finally, we demonstrate an initial implementation of a linked data-based recommender system by applying the concept of Linked Data Semantic Distances proposed by Passant (2010).

Pattern discovery and representation

First, data about existing buildings is retrieved and transformed into linked data. We hereby suggest to rely on the overall LBD approach documented earlier in Petrova et al. (2018*a*,*b*). This process is displayed on the bottom right in the system architecture diagram in Figure 4. For describing sensors, the LBD graph can be enriched with sensor node instances and sensors, as can be seen in Figure 2. Listing 1 lists all namespaces and prefixes used in the following examples.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix buildings: <https://www.example.com/data/buildings/> .
@prefix people: <https://www.example.com/data/people#> .
@prefix ls: <https://www.example.com/voc/linkset#> .
@prefix bmeta: <https://www.example.com/voc/buildingmetadata#>
    .
```

Listing 1: Namespaces and prefixes used in the following examples

As previously indicated, including sensor measurements can be done by pointing to an SQL store via a Web API or by including the sensor measurements explicitly in the graph. In this case, pattern discovery can be done using traditional data mining, which works with data batches and uses the previously discussed predictive and/or descriptive models. As explained in Petrova et al. (2018*a*,*b*), the resulting performance patterns can also be stored directly in the graph.

Alternatively, it is possible to convert the sensor data streams into RDF streams and perform semantic data mining on the resulting graph. Ideally, the RDF graph is first completed, which requires reasoning through the data and ontologies, and inferring all implicit data (e.g. `subclassOf` relations). To analyze how RDF stream processing would affect the recommendation concept, we employ the method described by Belghaouti et al. (2016), who identify frequent RDF patterns in RDF streams by mapping the graphs to adjacency matrices based on the graph predicates. Using this method, one is able to construct bit vectors, which describe the graph structure. Each bit vector is constructed from the predicates in the graph. The graph in Figure 2, for example, would lead to a bit vector (1111) that indicates each of the four predicates ('rdfs:label', 'gig:hasSensorNode', 'rdf:type', and 'fog:asObj- v3.0-obj'). All predicates and corresponding bit vector indices are recorded in a predicate hash table, which detects the patterns in the streams based on the bit vectors present in the graphs (e.g. 1111, 11101, 101, etc). Finally, a graph hash table is constructed, which records the frequency of occurrence of each bit vector. In this case, considering that all observations in the stream are modelled with the same predicates as in Figure 2, only one pattern would be included in the graph hash table, even though very diverse observation measurements are present. This has a big impact on pattern discovery, as the RDF frequent pattern mining is data structure oriented and considers the graph predicates instead of data values, as opposed to traditional data mining, which focuses only on data values.

User profiling and feedback

User profiling is a required feature for a well-functioning user-centred recommender system. We have set up the profiling system in a way similar to the one proposed in Boratto et al. (2017) (top of Figure 4). At user registration, a *Profile Initiator* component fills an RDF-based *User Profile Store*. These RDF-based profiles are built using the FOAF[9] ontology, and the result is an initial RDF graph identifying a user and its key metadata (Listing 2). The user is served recommendations through the *Recommendation Filter* component. All actions that the user takes in direct interaction with the recommender system are logged through a *Profile Learner* component. These actions serve as 'feedback' to the system, and they may come from a user clicking a 'like' button, a 'category' button, an 'annotation' button, or any other form of interaction. The Profile Learner component feeds back user profile data and user logs into the back-end of the recommendation system, which contains the *User Profile Store* and the *User Log Store*. In other words, the User Profile Store gets modified incrementally under the effect of the user interactions. The interactions of highest relevance are those related to recommendations, which are used by the end user in the project, especially if they respond directly to specific design requirements and performance targets.

```
people:EkaterinaPetrova
  a foaf:Person ;
  foaf:name "Ekaterina Petrova"^^xsd:string ;
  foaf:givenName "Ekaterina"^^xsd:string ;
  foaf:familyName "Petrova"^^xsd:string ;
  foaf:nick "epetrova"^^xsd:string .
```

Listing 2: People profile data

Feedback from user interaction goes into the User Logs and User Profiles, but the links between specific user profiles and relevant items in the *Building Data Store* are also kept, thereby aiming to enable a context-aware system. This means that we store links between user profiles and building identifiers in a separate RDF linkset (Listing 3).

[9] http://xmlns.com/foaf/spec/

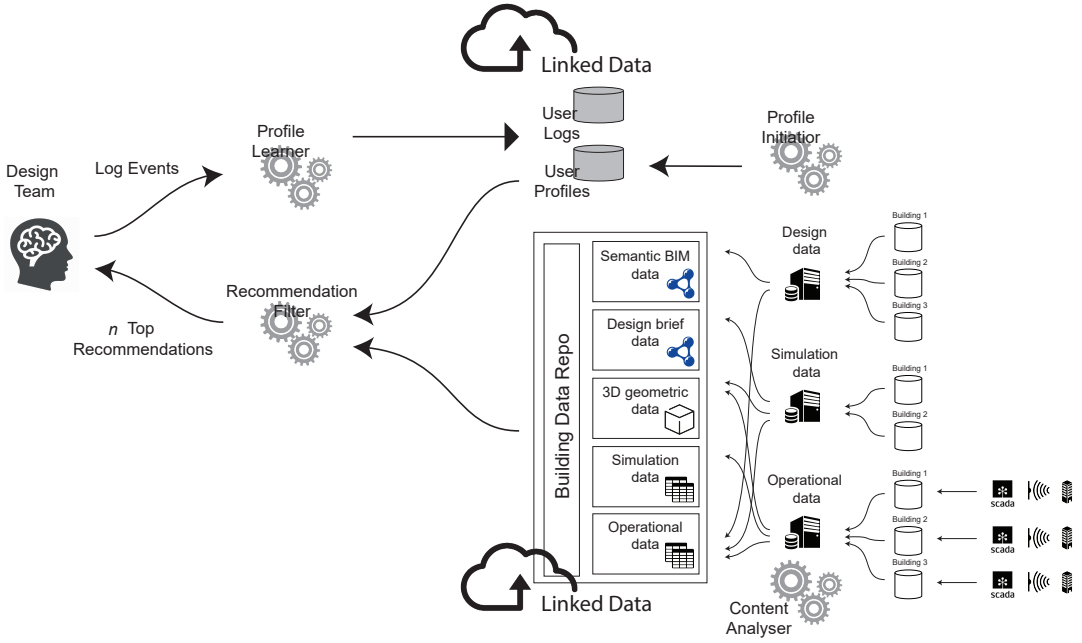

Figure 4: System Architecture for a linked data-based recommender system in the AEC industry.

This linkset serves as a hash table with identifiers from user profiles and the building data repository. Note that Listing 3 only includes `ls:like` relations, but other, more specific relations could be used as well, depending on how user interaction and feedback is tracked.

```
people:EkaterinaPetrova
  ls:likes buildings:building_987d706d-877a-4b1d-80f6-6
    ee89d856319 ;
  ls:likes buildings:building_af41d889-f50c-456e
    -9625-96655150838d .
```

Listing 3: Linkset between buildings and people.

We have applied this principle to the Building Data Store, User Profile Store, and Linkset Store as follows. Through user interaction and KDD, implicit data is retrieved about the buildings in the building data repository. As such, the buildings can be enriched with metadata tags. The result is displayed for two example buildings in Listing 4. Whereas this example only includes four simple metadata tags (`buildingType`, `designedBy`, `energyLabel`, `sustainabilityCertificate`), many more metadata tags can be used, e.g. category, occupancy data, mined performance patterns, design requirements, energy source, etc. These metadata can be used to form categories of design references, to compose queries in the database, to sort search results in a certain dimension, etc.

```
buildings:building_00dd6c87-6a6e-f482-7490-e6613659708a
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectX ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:LEEDPlatinum .

buildings:building_2e0dcc1c-b981-4c47-adb4-2b9887f10481
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectY ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:DGNBGold .
```

Listing 4: Example building data in TTL format.

In summary, the system holds three RDF-based data stores (besides the User Log Store): the User Profile Store, the Building Data Store, and the Linkset Store. It is now possible for an end user to query each of these stores for relevant data. For example, an end user may query for all buildings of a particular type, category and/or with a specific energy label (Listing 5). In this case the `bmeta` tags are used in the query. Of course, it is also possible to include user preference (Linkset Store) or user profile (User Profile Store) data in the queries. The returned results can be displayed to an end user, who is then able to sort the results using the available attributes and categories.

```
SELECT *
WHERE {
    ?b a bot:Building .
  ?b bmeta:buildingType bmeta:theater ;
    ?b bmeta:energyLabel bmeta:A .
}
```

Listing 5: Query for buildings of a particular building type.

Generating recommendations

As stated in the state of the art section, recommender systems can rely on the computation of semantic distance, i.e. the semantic relatedness between resources. Instead of limiting only to queries from within the end user environment, our recommender system should also make suggestions of buildings that are semantically close to, for example, a building that is considered to be most relevant to an end user at some point in time. Such buildings are the generated recommendations.

A set of measures were proposed in Passant (2010) to represent the 'Linked Data Semantic Distance' (*LDSD*) between two concepts (values between 0 and 1). This includes Direct, Indirect, and Combined Semantic Distance ($LDSD_d$, $LDSD_i$, $LDSD_c$), each either weighted or not. These semantic distances are used in recommender systems to find out what else users may like based on their user profile, search behavior, favorites, etc. The smaller the semantic distance between two related concepts, the higher the related concept is ranked in the set of n top related concepts or recommendations.

The semantic distance can be computed using all outgoing and incoming links of two concepts. For example, two different buildings might both be of type `theater`, which connects them to the same node for the `bmeta:buildingType` predicate, and makes them semantically closer. Determination of LDSD for recommendations starts as soon as an end user clicks a building from a result set that was previously returned with a simple query. In other words, the Recommendation Filter component is set up to look for 'bot:Building' objects that are semantically close to each other. The calculation hereby relies on all incoming and outgoing links for specific buildings, which are linked in the Building Data Store and the Linkset Store. Essentially, the simple indirect distance as a matrix between one building and all related buildings is calculated (Passant 2010).

This is illustrated in a simple example in Table 1, which shows the semantic distances for one of the buildings in the Building Data Store. As the `bot:Building` tag is present for all concepts, it is disregarded. Of course, in this limited example with 6 buildings and 3 relations (`buildingType`, `designedBy`, `energyLabel`), values are quite far apart (1/3, 1/2, 1, or 0), because only three links are considered. In an actual Building Data Store, semantic distances are much more interesting and diverse.

For each of the retrieved buildings, any available data can be displayed. This may also include sensor measurements and patterns found in them, metadata and user data in support of the end user, etc. Of course, this data needs to be displayed in an appropriate end-user interface, which is out of scope here.

Challenges and limitations

In terms of effectiveness of the proposed system, potential challenges need to be overcome, which can be related to user behaviour or the method that the recommendations are based on. Besides the knowledge base, the users play an important role in a recommender system. Important to consider are changes over time in the user profiles and preferences, which need to be taken into account continuously. End users may have similar profiles, but different behaviour and preferences depending on their context. These phenomena can affect the accuracy of a recommendation system, as the wrong user preferences may be considered. Anomalous behaviour such as disliking of recommendations also needs to be analysed and factored in. Another limitation may stem from the fact that despite being efficient, the LDSD approach only computes the semantic distance between two resources that are directly or indirectly linked through an intermediate resource. Therefore, enhanced LDSD algorithms may need to be used to expand the range beyond the two links distance. Also, the current system only considers semantic distances between buildings. Other semantic distances may be used as well, to configure and refine the recommendations.

Conclusions

Recent years have shown a rapid increase in technology uptake aiming to help reduce the negative environmental impact from buildings. This research attempts to minimise the negative contribution by informing the design with evidence-based feedback stemming from the existing building stock through a recommender system. Research on recommender systems has a long history, but is seldom implemented in the AEC industry. We attempt to overcome this challenge with data mining and linked data technologies. The article presents a state of the art review in the areas of KDD, semantic web technologies, stream processing and recommender systems. We investigate how to make sensor data streams efficiently available to the end user in addition to knowledge discovered in the data through semantic sensor technologies, web API connections, and/or sensor data stream processing. We outline the necessary steps towards a linked data-based recommender system, thereby drawing on the techniques that have been identified as most promising in the literature review. The software architecture of the proposed system consists of triple stores, as well as mechanisms for feedback handling and recommendations, data mining, and user interaction. Future work should focus on further implementation in practice including identifying how metadata tags can be inferred in the most intelligent way. Furthermore, the way sensor data are combined with semantic data, so that they can be used effectively in recommendation filtering, needs to be further investigated.

Table 1: Simple indirect semantic distances computed for https://www.example.com/data/buildings/building_00dd6c87-6a6e-f482-7490-e6613659708a.

Building	Cio	Cii	LDSD
https://www.example.com/data/buildings/building_2e0dcc1c-b981-4c47-adb4-2b9887f10481	2	0	0.3333
https://www.example.com/data/buildings/building_987d706d-877a-4b1d-80f6-6ee89d856319	1	0	0.5
https://www.example.com/data/buildings/building_43576e80-cf8c-11e1-8000-68a3c4d40f59	1	0	0.5
https://www.example.com/data/buildings/building_af41d889-f50c-456e-9625-96655150838d	0	0	1.0
https://www.example.com/data/buildings/building_aac3427f-eeb0-460c-ba47-14fd44c8be74	0	0	1.0

References

Barati, M., Bai, Q. & Liu, Q. (2016), SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data, *in* 'PRICAI 2016: Trends in Artificial Intelligence', Springer International Publishing, Cham, pp. 30–43.

Barbieri, D. F., Braga, D., Ceri, S., Della Valle, E. & Grossniklaus, M. (2010), 'C-SPARQL: A continuous query language for RDF data streams', *International Journal of Semantic Computing* **4**(1), 3–25.

Belghaouti, F., Bouzeghoub, A., Kazi-Aoul, Z. & Chiky, R. (2016), Fregrapad: Frequent rdf graph patterns detection for semantic data streams, *in* '2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)', pp. 1–9.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The Semantic Web', *Scientific American* **284**(5), 34–43.

Bizer, C., Heath, T. & Berners-Lee, T. (2009), 'Linked data - the story so far', *International Journal on Semantic Web and Information Systems* **5**(3), 122.

Boratto, L., Carta, S., Fenu, G. & Saia, R. (2017), 'Semantics-aware content-based recommender systems: Design and architecture guidelines', *Neurocomputing* **254**, 79–85.

Calbimonte, J.-P., Corcho, O. & Gray, A. J. G. (2010), Enabling ontology-based access to streaming data sources, *in* 'The Semantic Web – ISWC 2010', Springer, Berlin, Heidelberg, pp. 96–111.

Calbimonte, J.-P., Jeung, H., Corcho, O. & Aberer, K. (2012), 'Enabling query technologies for the semantic sensor web', *International Journal on Semantic Web and Information Systems* **8**(1), 43–63.

Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y. & Li, Y. (2016), Case studies of fault diagnosis and energy saving in buildings using data mining techniques, *in* 'Proceedings of IEEE international conference on automation science and engineering', IEEE, Fort Worth, TX, pp. 645–651.

de Wilde, P. (2014), 'The gap between predicted and measured energy performance of buildings: A framework for investigation', *Automation in Construction* **41**, 40 – 49.

Della Valle, E., Ceri, S., van Harmelen, F. & Fensel, D. (2009), 'It's a streaming world! reasoning upon rapidly changing information', *IEEE Intelligent Systems* **24**(6), 83–89.

D'Oca, S. & Hong, T. (2014), 'A data-mining approach to discover patterns of window opening and closing behavior in offices', *Building and Environment* **82**, 726–739.

Fan, C., Xiao, F., Madsen, H. & Wang, D. (2015), 'Temporal knowledge discovery in big bas data for building energy management', *Energy and Buildings* **109**, 75–89.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**(3), 37–54.

Han, J., Kamber, M. & Pei, J. (2012), *Data mining concepts and techniques*, Morgan Kaufmann, Waltham, US.

Hand, D., Mannila, H. & Smyth, P. (2011), *Principles of Data Mining*, MIT Press.

Heylighen, A., Martin, M. & Cavallin, H. (2007), 'Building stories revisited: Unlocking the knowledge capital of architectural practice', *Architectural Engineering and Design Management* **3**(1), 65–74.

Lausch, A., Schmidt, A. & Tischendorf, L. (2015), 'Data mining and linked open data new perspectives for data analysis in environmental research', *Ecological Modelling* **295**, 5–17.

Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I. & Novak, P. K. (2011), Using Ontologies in Semantic Data Mining with SEGS and g-SEGS, *in* 'Discovery Science', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 165–178.

Llanes, K., Casanova, M. & Lemus, N. (2016), 'From sensor data streams to linked streaming data: A survey

of main approaches', *Journal of Information and Data Management* **7**(2), 130–140.

McGlinn, K., Wagner, A., Bonsma, P., McNerney, L. & O'Sullivan, D. (2019), 'Interlinking geospatial and building geometry with existing and developing standards on the web', *Automation in Construction* . in press.

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S. & Rezgui, Y. (2017), 'Usability evaluation of a web-based tool for supporting holistic building energy management', *Automation in Construction* **84**, 154–165.

Middleton, S. E., Shadbolt, N. R. & De Roure, D. C. (2004), 'Ontological user profiling in recommender systems', *ACM Transactions on Information Systems (TOIS)* **22**(1), 54–87.

Miller, C., Nagy, Z. & Schlueter, A. (2015), 'Automated daily pattern filtering of measured building performance data', *Automation in Construction* **49**, 1–17.

Miller, C., Nagy, Z. & Schlueter, A. (2018), 'A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings', *Renewable and Sustainable Energy Reviews* **81**, 1365 – 1377.

Musto, C., Basile, P., Lops, P., de Gemmis, M. & Semeraro, G. (2017), 'Introducing linked open data in graph-based recommender systems', *Information Processing and Management* **53**, 405–435.

Nebot, V. & Berlanga, R. (2012), 'Finding association rules in semantic web data', *Knowledge-Based Systems* **25**(1), 51–62.

Oliveira, J., Delgado, C. & Assaife, A. (2017), 'A recommendation approach for consuming linked open data', *Expert Systems With Applications* **72**, 407–420.

Passant, A. (2010), Measuring semantic distance on linking data and using it for resources recommendations, *in* 'AAAI spring symposium 2010: Linked data meets artificial intelligence', pp. 93–98.

Pauwels, P. (2014), 'Supporting decision-making in the building life-cycle using linked building data', *Buildings* **3**, 549–579. DOI: 10.3390/buildings4030549.

Pauwels, P. & Terkaj, W. (2016), 'EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology', *Automation in Construction* **63**, 100–133.

Pauwels, P., Zhang, S. & Lee, Y.-C. (2017), 'Semantic web technologies in aec industry: a literature review', *Automation in Construction* **73**, 145–165.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*a*), From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches, *in* 'Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM)', pp. 391–398.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*b*), In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, *in* 'Proceedings of the CIB W78 Conference', pp. 19–26.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*c*), 'Towards data-driven holistic sustainable design: A decision support framework relying on knowledge discovery in real-time building performance data and disparate project data repositories', *Architectural Engineering and Design Management* pp. 1–23.

Rasmussen, M. H., Pauwels, P., Hviid, C. A. & Karlshøj, J. (2017), Proposing a central AEC ontology that allows for domain specific extensions, *in* 'Proceedings of the Joint Conference on Computing in Construction'.

Ristoski, P. & Paulheim, H. (2016), 'Semantic web in data mining and knowledge discovery: A comprehensive survey', *Web Semantics: Science, Services and Agents on the World Wide Web* **36**, 1–22.

Sacks, R., Lee, C. M. E. G. & Teicholz, P. (2018), *BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators*, 3 edn, John Wiley & Sons, Hoboken, NJ, USA.

Schneider, G. F. (2017), Towards aligning domain ontologies with the Building Topology Ontology, *in* '5th Linked Data in Architecture and Construction Workshop'.

Shekhar, S., Zhang, P. & Huang, Y. (2010), Spatial data mining, *in* 'Data Mining and Knowledge Discovery Handbook 837-854.', Springer.

Sheth, A., Henson, C. & Sahoo, S. (2008), 'Semantic sensor web', *IEEE Internet Computing* **12**(4), 78–83.

Wang, Z. & Srinivasan, R. (2017), 'A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models', *Renewable and Sustainable Energy Reviews* **75**, 796–808.

Xiao, F. & Fan, C. (2014), 'Data mining in building automation system for improving building operational performance', *Energy and Buildings* **75**, 109–118.