Semantic Framework for Facilitating Product Discovery

Paras Bhutani, Shubham Kumar Baranwal and Sarika Jain

National Institute of Technology, Kurukshetra, India

Abstract

With the rapid growth in the field of e-commerce, online shopping has become a major part in lives of people. Online stores provide search engines for product discovery. Traditional search engines are working on syntactic approach, because of which these search engines are not able to extract precise result. Since products catalogues comes from different sources, there comes need of binding the information of these product catalogues into a single taxonomy for better categorization of products. By using semantic web technologies, we can represent all the information in machine understandable form. Semantic search engine use knowledgebase to store and retrieve the information. Knowledgebase uses a semantic model that define all the relationships, classes and properties. Instances are annotated according to this model and a knowledge graph is created. This intelligent binding of data helps the semantic search engine to understands the user query and provides accurate result. We propose a semantic search engine framework for facilitating product discovery which uses an ontology for product categorization and annotate all the product catalogues from different retailers into a single knowledge graph or RDF. After that the RDF store is used for storing and retrieving data using semantic queries.

Keywords

E-Commerce, Knowledge Graph, Ontology, Product Discovery, Semantic Search, RDF

1. Introduction

Online shopping is comparatively easier and better than traditional shopping because nowadays everything is available to us at our door steps just what we need to have is a good availability of internet. An inefficient product searching has been a major cause of decrements in sales, continuous poor service [1,3,8] can affect the business likely for evoking new customers as well. As stated by [15] Statista Statistic report in 2018, an estimated 1.8 billion people around the world buy goods online. During the same year, global e-retail sales amounted to 2.8 trillion U.S. dollars and projections show a growth of up to 4.8 trillion U.S. dollars by 20.

The Fig. 1 [13] shows the snags which online customers face when they search for products on e-Commerce websites. As per the statistics[14] published on Marketing Charts.com in June 2018, Data Source: RichRelevance, many customers end up with irrelevant product results (28 %), unable to find the product they need (24 %) , search function does not recognize the words used by particular customer(18 %), unclear search box(10 %) and many more. The absence of HTML standard to introduce the heterogeneous e-commerce information which results in the low quality of web crawler results, which makes clients to invest time and energy for arranging and comparing items between sites and picking the correct item that suits their

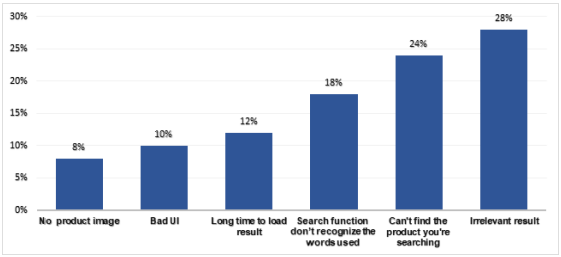

Figure 1: Shoppers'Frustration with Retail Site Search.

requirements [4,20]. There are certain problems with the traditional approaches. Interoperability is also a major problem that arises due to the heterogeneous formats of information representation[11]. Because of innovation new features are added to products frequently which requires continuous changes in schema. The basic problem in keyword-based search is it only finds keywords of user query on the web and list all the pages in result[18]. For example- if you are searching on a keyword-based search engine "Books about hotel", this will display you the web pages about “hotels”, “books”, “hotel booking” and “books about hotels” but actually you were searching for the books written about hotels. The keyword-based search gives better hit ratio because it works on two keywords hotel and books but semantic precision and recall is low in this. We want it to understand the intent of user but it only understands the keyword from the web pages.

Web technologies are changing rapidly from static web to progressive web, then to semantic web [23]. Semantic web helps machine to interpret the meaning of data. Semantic web technologies can be applied in various fields

ACI'21: Workshop on Advances in Computational Intelligence at ISIC 2021, February 25-27, 2021, Delhi, India

parasbhutani13@gmail.com (P. Bhutani);
baranwalshubham19@gmail.com (S. K. Baranwal);
jasarika@nitkkr.ac.in (S. Jain)
0000-0002-7432-8506 (S. Jain)

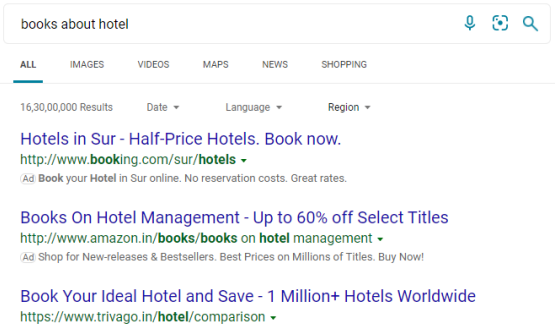

Figure 2: Result on Keyword-based Search Engine.

where data integration improves the quality like oil and gas industry, pharma industry, movies/series searching. In all these fields a huge amount of data is generated every day. Semantic search engine has its own benefits which results comes in relevant matches rather than providing unnecessary product information. As everything is defined using relationship it's easy to understand customer's requirements which results in better relationship of enterprise-customer. If product sites documents are designed semantically, one may get better results while searching and surely that will profitable to the business.

For online business product database is the wealth of e-business and the site is its appearance, the "virtual" exhibit for clients. For obtaining this we have used the knowledge base to store the data effectively. Knowledge base work like human brain to organize information in an abstract way. In knowledge-base a semantic model is present to better classify the data using classes, subclasses, relationships and instances. Ontology has the ability to define such model. The goal is to create a semantic model for sharing information about a specific domain. There are various product classification standards which exists. These help for better online data handling, dealing with business related terms and services. Few of them are eClassOWL , GoodRelations , Global Product Classification, Schema.org etc. We have also developed an ontology for categorization of electronic devices. The next step data integration involves knowledge graph that represents the information in RDF is used for data integration. Benefit of using knowledge graph is that it provides expressivity, performance and standardization. A triple store or RDF store is used for the storage and retrieval of information from RDF using SPARQL queries with the help of SPARQL endpoint.The key idea here is to semantically annotate all the products catalogues according to a single taxonomy and store it into triple store. Our contribution is mainly in two directions i.e., we have created an ontology for product categorization. and a working prototype using semantic technologies for facilitating product discovery.

2. Related Work

There are two ways for product categorization one is using machine learning techniques and another one is using lexicon-based techniques[2]. As our work is on semantic web technologies so we are discussing here work related to semantic web technologies. The authors in [5] have described about product information retrieval framework that uses OA-VSM and SPARQL-based approach rather than VSM-based keyword search. Ontology based adaption of vector space model retrieval method helps the user to get better result. The authors in [6] have introduced the importance of availability of machine understandable information. Machine understandable information has the potential to impact many significant web applications that includes searching. Two mechanisms are discussed that have increasing search results from the semantic web. Angelo A. Salatino et al. [7] presented CSO classifier as a tool which classifies text according to the Computer Science Ontology. CSO Classifier takes input in form of unstructured text, title, and keywords of a research paper and classify the research area of paper as output. E. Peis et al [9] states that recommender systems based on semantic technologies give out high performance because they are based on a knowledge base mostly defined by an ontology. Using ontology solve interoperability problem, improve performance in social networks, represent information semantically and improve searching result.

The authors in [16] stated the importance semantic web mining which means that semantic adds meaning to the data but web mining use to filter out that data more accurately. The basic idea is to extract the useful information from the bulk and for that we need standardized ontologies. Example- Goodrelations, e-class OWL. Kerie A. [17] stated the current limitation of searching on the web and the reason of these problems are syntactic and semi semantic approach. On the basis of ontology, they have tried to correct these problems. Here they have made EEPS (Ethiopian Export Products and Services Domain) ontology which is consistent, flexible and moderate in size which makes the management of the information better. They study [19] the challenges that traditional e-commerce faces and overview of how semantic web technologies overcome these challenges in e-commerce field. They have developed a semantic search engine for e-commerce field using GOODRELATION ontologies. The authors in [21] designed a framework based on ontology for the purpose of search and retrieval of products. They find out that the ontology created for searching performs better for searching purpose rather than catalog side ontologies. Necula S. [22] in their study find out that customer consider knowledge graph is important for an

e-commerce website. Another finding of their study is that searching a product by categories and subcategories gives better result.

3. Proposed Framework

To overcome the problem of irrelevant search of products on online store our work uses a semantic search engine. The aim of our work is to provide customer best product according to their requirements. Figure 3 presents a general architecture of our proposed framework. The proposed system has two interfaces, one for retailers in order to store new products and another one for customers to get recommendation according to their input specification about products. We have divided proposed framework into two modules. In first module i.e., in knowledge modeling we will create schema for knowledgebase and after that instances are populated. All these data are then stored in triple store. After that in second module i.e., in knowledge consumption model end user enters the specification about products and a SPARQL query is processed over knowledge base and desired results are fetched according to the end users' requirements.

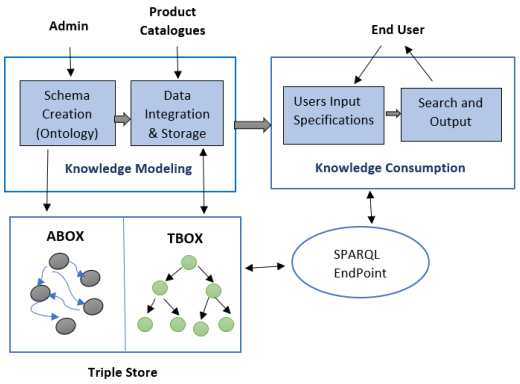

Figure 3: Proposed Architecture.

4. Materials and Methods

In this section we discuss about various tools and technologies that can be used to implement the proposed framework

4.1. Existing Product Classification Standards

There are various product classification standards which exists. Out of these some of them are mentioned below which are useful for us.

- **eClassOWL:** eClassOWL released in 2004. eClassOWL is intended to be utilized for product classification. It has 30,000 classes and 5,000 properties of product features.
- **GoodRelation:** GoodRelations is a vocabulary used for sharing e-commerce information. It has 30,000 classes and 15 properties of product features. Now it is integrated into schema.org and supported by Google, Yahoo etc.
- **Global Product Classification:** GPC is utilized for grouping all the products into a common language. The building block of GPC is an item code known as a brick. There are bricks for everything from a vehicle to a bottle of milk. The most elevated level of the characterization is a fragment, which is characterized as a specific industry. For instance, a bottle of milk has a place with the food, beverages and tobacco section.
- **UNSPSC:** UNSPSC is a vocabulary about products and services used in e-commerce field. It has 20,792 classes. UNSPSC empowers expenditure analysis at gathering levels pertinent to your requirements. For our proposed framework we have created our own product classification ontology.

For our proposed framework we have created our own product classification ontology.

4.2. Triple Store

This list shows notable triple store, APIs, and other storages that have implemented the W3C SPARQL standard.

- **Apache Jena:** Jena supports all operating system like Linux, Windows etc., with a Java Virtual Machine. It is an open-source software. It gives us an API to take out data from and write to RDF file. Apache Jena is completely FREE as open-sourced Java-based software from Apache Software Foundation. It includes a majority of semantic web technology standards, such as RDF API, SPARQL query language, and Ontology API.
- **Allegro graph:** It was created to satisfy World Wide Web Consortium guidelines for the RDF, so it is appropriately viewed as a Relational Database Framework. It is a reference usage for the SPARQL rules. These rules are a standard query language for connected information, filling similar needs for RDF databases that SQL serves for relational database. Allegro graph is a closed source triple store which is intended to store RDF.

Out of these listed tools we are using Jena because apache Jena supports all operating system like Linux, Windows

etc., with a Java Virtual Machine. It is an open source providing different API's to extract data from and write to RDF graphs. The graphs represented as an abstract model. A model can be sourced with data from files, Databases, URLs or combination of these.

4.3. RDFizer

There are various RDFizer tools used for converting the csv, json and other formats data into RDF. We are specifying different tools here:

- **EasyRdf:** It is an online tool which takes inputs in different formats and converts it into RDF format. It is quite easy to use even you can put your URI's to be fetched and converted into RDF forms.
- **CSV2RDF:** CSV2RDF is a java-based tool that is used for converting CSV files into generic triplets or RDF. For achieving this it uses CONSTRUCT query to create mappings.

Out of listed above we are using CSV2RDF because it is easy to use and we are taking product catalogues in CSV format in our proposed framework. It just takes CSV files as input and converts them into RDF format data. It is a java-based application and we also have implemented our framework in java hence we have used CSV2RDF tool

4.4. Integrated Development Environment

Many IDE are used in computer programming. They contain a base workspace and an extensible plug-in system for managing the environment. Few of them we are listing here:.

- **Eclipse:** It is broadly utilized among business organizations to make amazing applications for programming improvement, yet for different enterprises, for example, banking, automobiles, clinical and space investigation. Since it is coded in Java, Eclipse underpins most stages and working frameworks. Nonetheless, Eclipse isn't constrained to Java. It also allows us for using several languages by using their components of plugins.
- **NetBeans:** It is an integrated development environment (IDE) for Java. It allows us to create applications from a lot of small units of programs segments called modules. It supports various extensions of other programming languages like PHP, C, C++, HTML5, and JavaScript. This is an open-source integrated development environment.
- **IntelliJ:** It is an IDE coded in Java for creating computer programming. It is developed by JetBrains and is accessible as an Apache 2 Licensed people group edition. The IDE provides mix with manufacture/bundling devices. IntelliJ has a lot of features and it is designed for making large projects. It supports languages like Kotlin, Java script, SQL etc.

Out of these listed IDE we are using IntelliJ because we have created many projects using IntelliJ so we preferred IntelliJ for this task too and since apache Jena could be embedded easily so we opted the same. Setting up the environment in java using IntelliJ is quite easier. IntelliJ provides direct making of files of class java in it. IntelliJ can be compared to Android Studio because it is also used for making applications at the same time. It is also drag and drop desktop application software making application which allows you to create good GUI based application.

5. Implementation

In this section we are explaining steps followed while implementing each module of proposed framework to facilitate product searching. As we already discuss we have divided the task into two modules, knowledge modeling and knowledge consumption.

5.1. Knowledge Modelling

If you want to fetch a correct and accurate information according to your need than your knowledge base i.e., where the information is stored should be managed properly. We are using the domain ontology for storing our information. The ontology is structured vocabulary that defines the knowledge as concept[12]. Ontologies are framework which is used as bases for linked data modeling because of its ability to explaining relationships and interconnectedness between them. More specific knowledge can be captured because of the semantic relations of instances with their attributes[10]. This module further has two components: schema creation and data integration.

- **Schema Creation:** First step is to determine the scope and purpose of domain ontology. At this stage we determine the scope and purpose of creating ontology. Before making ontology our purpose of making it should clear. Our ontology based on electronic products provide us some semantic information about the products that improves product searching results. Therefore, the more semantic information about the product should be provided in order to have more and more accurate information. Second step is collecting information about domain. When the idea about domain information and the domain knowledge is collected,

now we are able to create a good ontology which describes our requirement correctly. Third step is defining the classes and topology. There are three methods to design the classes. First one is top-down method- In this first we create general class, then we create specific classes as per use. Second one is bottom-up method - In this first we create specific classes and then we create general classes. Third is the combination method - In this we create most important classes first and then move towards general classes. Last step of schema creation is defining the properties- only class cannot give all the information therefore we have to define the properties of classes. We should define properties of general classes because all its properties are inherited by its sub-classes.

We created an ontology about computers and laptops where we have defined following classes, object properties, data properties and structure of ontology as shown in images below:

Figure 4: Classes, Object properties, Data properties of our Ontology.

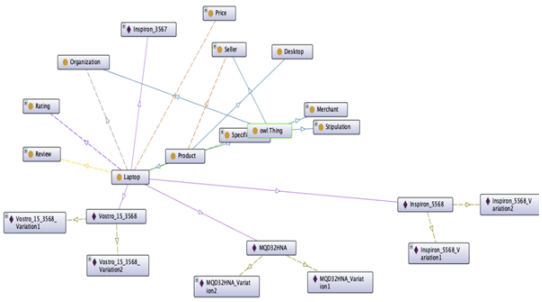

Figure 5: Graphical View of Ontology.

- **Data Integration:** At last we will add instances according to the schema we have created. For populating we first select a class then we create an instance for that and after we populate it as an instance to the category it belongs. When we populate instances on our knowledge model a knowledge graph is created. We can convert any format of data like csv, json, json-ld etc, according to our schema using RDFizer.

Initially we added seven instances i.e. details of five laptops/desktop manually using protege. Therefore, we have seven instances in our knowledge graph. We

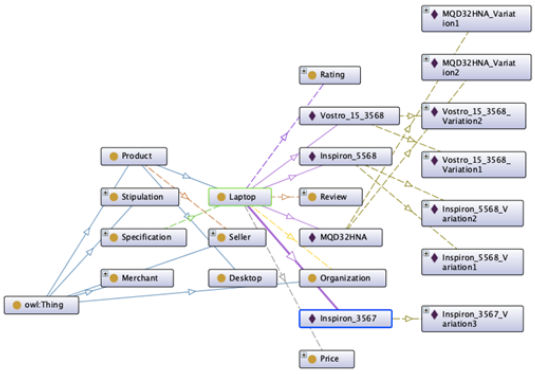

Figure 6: Manually Added Instances.

have created an interface to add instances directly to our knowledge graph i.e. in our RDF file. Retailer will upload laptop/computer catalogues in CSV format, thereby storing product catalogues in our knowledge-base for consumption (query) by the end user. Once

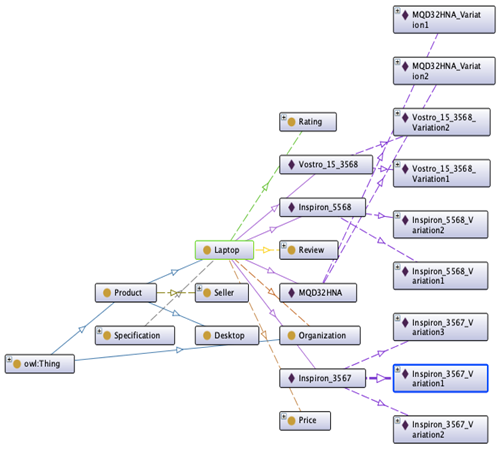

Figure 7: Updated Knowledge Graph.

we receive the catalogues, CSV2RDF a java-based application which use Jena API will convert it from CSV to RDF format by creating mapping between base URI of our ontology and CSV column. There are two scenarios, first is that we create separate RDF file for data provided by each retailer and after that we merge all the files into a single RDF file and second being that all the data can be annotated into a single RDF file

as per the requirements of the application. We have annotated all the data to a single RDF file.

Every time retailer uploads their product catalogues it is annotated semantically according to data model and all he instances added into knowledge graph. There are few new instances are present in updated knowledge graph shown in Figure 7.

5.2. Knowledge Consumption

In knowledge modeling module every information about product catalogues are stored in machine understandable form in triple store. In knowledge consumption module we query the information present in triple store and fetch out the products that matches with user requirements.

RDF files generated after data integration module are stored in a triple store, Jena tdb. Jena tdb provides APIs to perform SPARQL Queries.

An interface is created for end users to enter their specifications. On clicking search button SPARQL endpoint will generate a SPARQL query shown in Figure 8. which execute on knowledge graph and products instances matches with these specifications are fetched.

```
String queryString =
"PREFIX my:<http://purl.org/laptop-sa#>"+
    "PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>"+
    "SELECT ?x WHERE {" +
    "?Laptop my:RAM \""+ ram +"\"^^xsd:string."+
    "?Laptop my:Processor \""+ processor +
    "\"^^xsd:string." +
    "?Laptop my:Name ?x ."+
 s"}";
```

Figure 8: SPARQL Query.

For example user is looking for a laptop that has RAM = 4GB and has PROCESSOR = Intel Core i3. Two instances of laptops that are present in knowledgebase matches with these specifications i.e. Dell Vostro 15 3568 and Dell Inspiron 5568(Z564304SIN9) 2 In 1 laptop are recommended to end user.

6. Conclusion and Future Work

As we know the success of any e-commerce store depends on the database that how they have stored the information and how efficiently from it data can be retrieved. Products catalogues are in heterogeneous format. To organize information in a better way this work uses a knowledgebase which contains a semantic model that store information in a common format and in machine readable form. The challenge of efficiently using the information from the bulk of stored data has been made possible with the help of semantic web. The model helps us to get better result because semantic relationship helps the search engine to understand intent of user query in the same manner as human brain and hence products match with user requirements are fetched. In future we will evolve our ontology to add more classes about other electronic devices and publish our ontology on web and for input we are going to use web pages from which data is linked to our model and the instances will populate be populated into knowledge graph.

References

[1] Ristoski, Petar, Petar Petrovski, Peter Mika, and Heiko Paulheim. "A machine learning approach for product matching and categorization." Semantic web 9, no. 5 (2018): 707-728.

[2] Rif'at, Muhammad, Rahmad Mahendra, Indra Budi, and Haryo Akbarianto Wibowo. "Towards product attributes extraction in Indonesian e-commerce platform." Computación y Sistemas 22, no. 4 (2018).

[3] Shah, Kashif, Selcuk Kopru, and Jean David Ruvini. "Neural network based extreme classification and similarity models for product matching." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). 2018.

[4] Al-Taie, Mohammed Zuhair, Siti Mariyam Shamsuddin, and Joel Pinho Lucas. "Predicting the Relevance of Search Results for E-Commerce Systems." Int. J. Advance Soft Compu. Appl 7.3 (2015).

[5] Zhang, Liyi, Mingzhu Zhu, and Wei Huang. "A framework for an ontology-based e-commerce product information retrieval system." Journal of Computers 4.6 (2009): 436-443.

[6] Guha, R., R. McCool, and E. Miller. "Semantic search. WWW'03: Proceedings of the 12th international conference on World Wide Web." (2003): 700-709.

[7] SELECTUSA retailpage, https://www.selectusa.gov/retail-services-industry-united-states

[8] Salatino, Angelo, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. "Classifying research papers with the computer science ontology." (2018).

[9] Peis, Eduardo, JM Morales del Castillo, and Juan A. Delgado-López. "Semantic recommender systems. analysis of the state of the topic." Hipertext. net 6.2008 (2008): 1-5.

[10] Kutiyanawala, Aliasgar, and Prateek Verma. "Towards a simplified ontology for better e-commerce search." arXiv preprint arXiv:1807.02039 (2018).

[11] Jain, Sarika. "Exploiting Knowledge Graphs for Facilitating Product/Service Discovery." arXiv preprint arXiv:2010.05213 (2020).

[12] Ontotext, https://www.Ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

[13] BrainSINS, https://www.brainsins.com/en/blog/ai-powered-semantic-search-increase-ecommerce/conversion/4526

[14] Marketing charts,https://www.marketingcharts.com/digital/seo-83805

[15] Statista https://www.statista.com/topics/871/online/shopping/

[16] Aslam, Zulqurnan. "Semantic Web Mining in E-Commerce Websites." International Journal of Computer Applications 137.2 (2016).

[17] Minch, Arba. "Engineering Semantic Web for E-Commerce Business Intelligence: a Bilingual EEPS Ontology Model."PhD diss., Arba Minch University, 2014.

[18] Malve, Ankita, and P. P. M. Chawan. "A comparative study of keyword and semantic based search engine." International Journal of Innovative Research in Science, Engineering and Technology 4.11 (2015): 11156-11161.

[19] Benjamins, Victor Richard, Jesús Contreras, Oscar Corcho, and Asuncion Gomez-Perez. "The six challenges of the Semantic Web." (2002)..

[20] Sonika Malik, Sarika Jain, "Semantic Ontology-Based Approach to Enhance Text Classification", CEUR Workshop Proceedings, vol 2786, pp. 85-98. International Semantic Intelligence Conference, Delhi, India. 25-27 Feb 2021.

[21] Kutiyanawala, Aliasgar, and Prateek Verma. "Towards a simplified ontology for better e-commerce search." arXiv preprint arXiv:1807.02039 (2018).

[22] Necula, Sabina-Cristiana, Vasile-Daniel Păvăloaia, Cătălin Strîmbei, and Octavian Dospinescu. "Enhancement of e-commerce websites with semantic web technologies." Sustainability 10, no. 6 (2018): 1955.

[23] Rudman, Riaan, and Rikus Bruwer. "Defining Web 3.0: opportunities and challenges." The Electronic Library (2016).

[24] Jain, Sarika. Understanding Semantics-Based Decision Support. CRC Press, 2020.