An Open Source Toolkit to Parse and Analyze Online Crowdsourced Portals

Amit Arjun Verma (2016csz0003@iitrpr.ac.in)
Indian Institute of Technology Ropar https://orcid.org/0000-0002-2392-0198
S.R.S Iyengar
Indian Institute of Technology Ropar
Simran Setia
Indian Institute of Technology Ropar
Neeru Dubey
Indian Institute of Technology Ropar

Research

Keywords: Knowledge Building, Wikipedia, Stack Exchange, open-source, Python library

Posted Date: January 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-154658/v1

Abstract

With the success of crowdsourced portals, such as Wikipedia, Stack Overflow, Quora, and GitHub, a class of researchers is driven towards understanding the dynamics of knowledge building on these portals. Even though collaborative knowledge building portals are known to be better than expert-driven knowledge repositories, limited research has been performed to understand the knowledge building dynamics in the former. This is mainly due to two reasons; first, unavailability of the standard data representation format, second, lack of proper tools and libraries to analyze the knowledge building dynamics.

We describe Knowledge Data Analysis and Processing Platform (KDAP), a programming toolkit that is easy to use and provides high-level operations for analysis of knowledge data. We propose Knowledge Markup Language (Knol-ML), a generic representation format for the data of collaborative knowledge building portals. KDAP can process the massive data of crowdsourced portals like Wikipedia and Stack Overflow efficiently. As a part of this toolkit, a data-dump of various collaborative knowledge building portals is published in Knol-ML format. The combination of Knol-ML and the proposed open-source library will help the knowledge building community to perform benchmark analysis.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

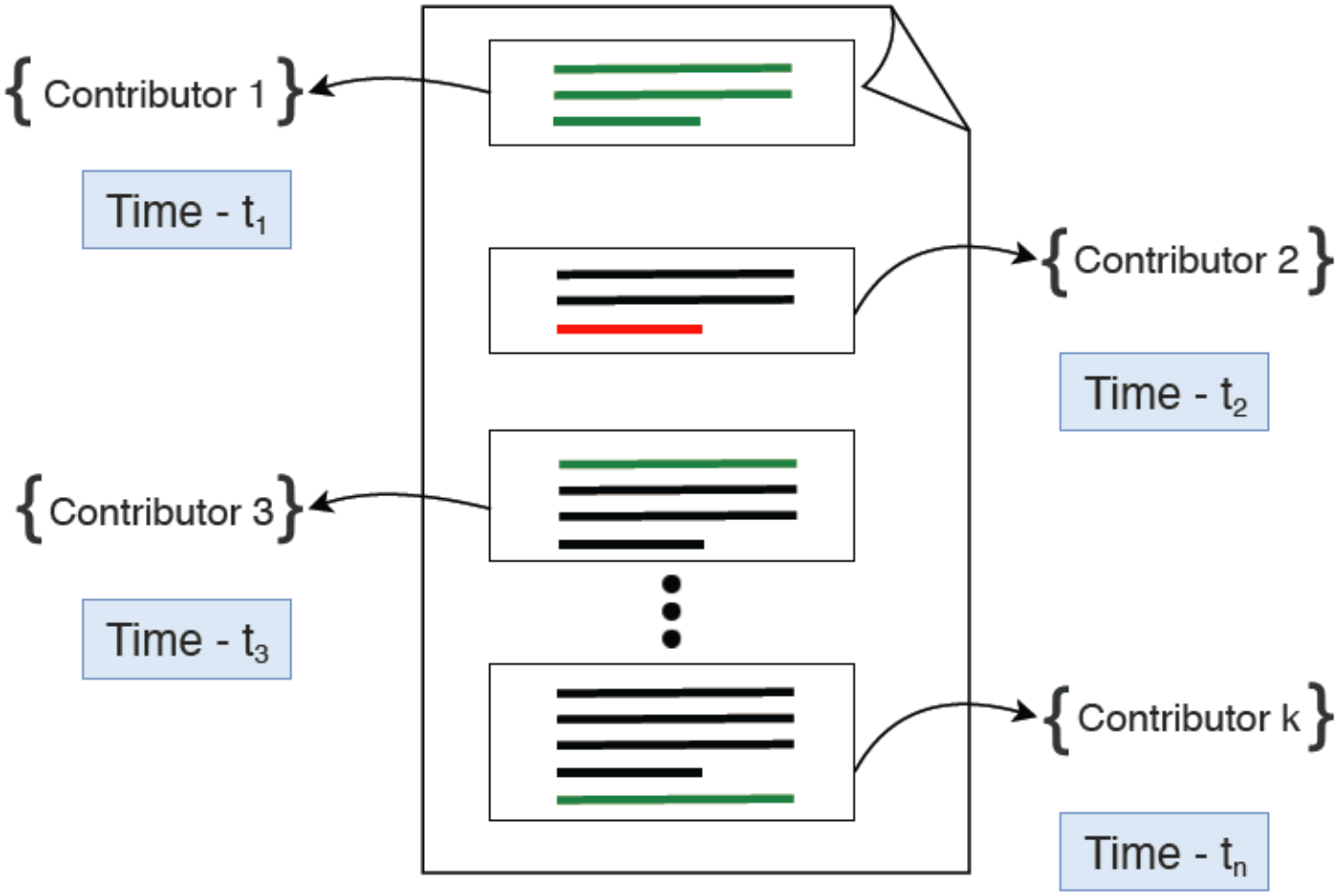

Figure 1

An illustration of sequential knowledge building in crowdsourced environ- ment. A contribution can be seen in the form of addition (green) or deletion (red) information.

```
<KnowledgeData Type="text" Id="1">
    <Title>This is an example for sequential knowledge data</Title>

    <Instance Id="1" InstanceType="Revision">
        <TimeStamp><CreationDate>t1</CreationDate></TimeStamp>
        <Contributors>
            <OwnerUserName>Editor 1</OwnerUserName>
            <OwnerUserId>1</OwnerUserId>
        </Contributors>
        <Body><Text Type="text">Edits of Editor 1</Text></Body>
    </Instance>

    <Instance Id="2" InstanceType="Revision">
        <TimeStamp><CreationDate>t2</CreationDate></TimeStamp>
        <Contributors>
            <OwnerUserName>Editor 2</OwnerUserName>
            <OwnerUserId>2</OwnerUserId>
        </Contributors>
        <Body><Text Type="text">Edits of Editor 2</Text></Body>
    </Instance>
        ⋮
</KnowledgeData>
```

Figure 2

Representation of sequential knowledge data in Knol-ML format.

```
<Def attr.name="abc" attr.type="string" for="Instance" id="abc" />

<KnowledgeData Type="text" Id="1">
    <Title>Title of the article</Title>

    <Instance Id="1" InstanceType="Revision">
        <TimeStamp><CreationDate>t1</CreationDate></TimeStamp>
        <Contributors>
            <OwnerUserName>Editor 1</OwnerUserName>
            <OwnerUserId>1</OwnerUserId>
        </Contributors>
        <EditDetails>
            <EditType>details of edit</EditType>
        </EditDetails>
        <Body><Text Type="text">Edits of Editor 1</Text></Body>

        <Knowl Def="abc">sample value</Knowl>

    </Instance>

        ⋮

</KnowledgeData>
```

Additional Data

Figure 3

Representation of extension mechanism in Knol-ML.

```
<KnowledgeData Type="text" Id="1">
    <Title> Question Title </Title>

    <Instance Id="1" InstanceType="Question">
        <TimeStamp>
            <CreationDate> t1 </CreationDate>
        </TimeStamp>
        <Contributors>
            <OwnerUserId> Editor 1 </OwnerUserId>
        </Contributors>
        <Body><Text Type="text"> Question </Text></Body>
        <Credit><Score> score value </Score></Credit>
    </Instance>
        ⋮
</KnowledgeData>
```

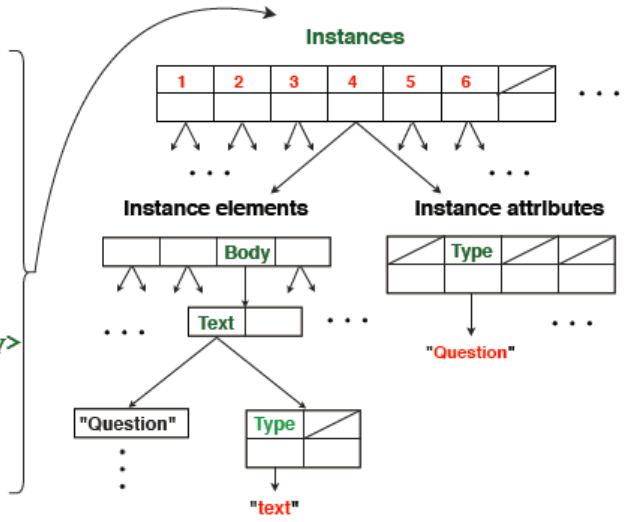

Figure 4

A diagram of knowledge data structures. Instance ids are stored in a hash table, and each instance has one hash table and one vector associated with it representing the elements of the instance.

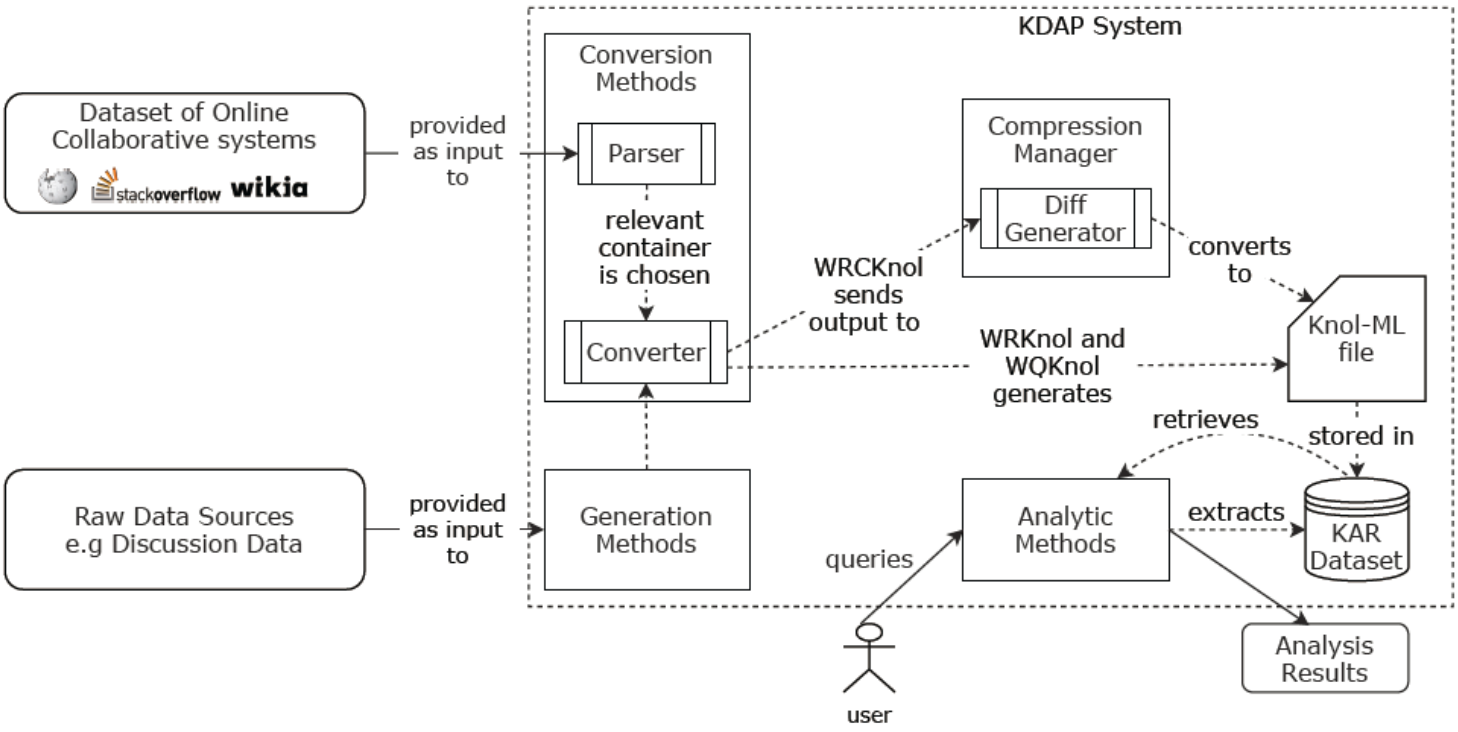

Figure 5

Overview of KDAP framework.