ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science & Information Technology (NCETCSIT-2011)

Study and Population of Artifacts

Chandni Singh, Rishabh Srivastava

***Abstract*— Artifacts are man-made objects taken as a whole. This paper presents various kinds of artifacts based on different division criteria and methods to create a list of artifacts. Different methods have been discussed and then we show how some of them can be used on specific kinds of text to create an exhaustive list of artifacts.**

***Index Terms*—artifacts, component, population, surface text patterns**

I. Introduction

Artifacts are man-made objects. Ontologies like WordNet, ConceptNet, PurposeNet, etc. often require a list of artifacts. A problem often faced in various domains is the shortage of a thorough artifact list. WordNet provides a list of only 10000 artifacts when there is estimated to be around 1 million artifacts in the world. In order to increase the number of artifacts we opted for many ways. Some of them included: -

- Manual population using Wikipedia titles.
- Population using lists available on the internet.
- Population using the artifacts available in Wikipedia.
- Population of the artifacts available in the dictionary.

II. Related Work

Though the methods discussed in this paper have been used in other fields, no previous work has been done for the population of a list of artifacts. Reference [5] gives an algorithm for identifying effective surface text patterns based on the frequency of occurrence of the patterns in the given text. References [6][7] have already shown effective use of STPs in extracting knowledge from large corpus. Extraction from large web corpora is explained in [4]. How the presence of 'defining formulae' in dictionary improves understanding of semantics of words has been clearly shown in [1].

III. Artifacts

Before we start with the population of a list of artifacts, we need to understand thoroughly what all can be termed as an artifact. An artifact is any man-made entity taken as a whole. A chair is an artifact; even a leg on which it stands is an artifact. Artifacts may be physical or abstract. A physical artifact is an entity that has a physical existence e.g. football, classroom, etc. An abstract artifact however is an entity that exists only abstractly e.g. team, school, etc. Physical artifacts include simple entities like paper, and even complex assemblies like car, television etc.

A territory is a region ma rked off for a certain purpose and hence an enclosed boundary and so a physical artifact while a kingdom is defined as a domain in which something is dominant, e.g. a kingdom of reason and hence an abstract artifact.There may be several different kinds of things to be attributed to as artifacts. Artifacts may be studied based on the following points:

- Artifacts may be simple. Simple artifacts have an independent existence. They are complete in themselves without the existence of any other artifact. E.g. a brick or nut has an independent existence.
- There may be complex artifacts that are not a single whole but formed by the assembly of various other artifacts. A bus is an artifact in itself, but is an assembly of various other complexes (e.g. engine which again is composed of many artifacts) and simple artifacts (e.g. front pane).
- Most artifacts today need to be prepared before they can be used directly. The preparation of artifacts may include providing energy from some power source like electricity, some refillable material like fuel in a bus, air in a tyre or ink in a pen or it may include some kind of change brought about by certain other entities like cooking of food by heat or by fire.
- Some artifacts may have a number of meanings as to the purpose they serve. The same motor might be used in a mixer as well as a grinder with a change in the amount of power that it generates and hence may need to be written as two separate artifacts i.e. mixer-motor and grinder-motor.
- Artifacts lying in the sub tree of other artifacts have specialized forms, which have almost the same set of features and physiology as that of a base type along with additional features. Variations in the properties of these artifacts are of significance from the inheritance point of view
- Two artifacts may have the same underlying principle but might have a big difference in the sizes of its components. A flour mill and a mixer used in households both work on the principle of rotation of a blade but have quite a difference in the sizes of their components and even the purpose that they serve.
- Artifacts sometimes have certain special components so as to aid in serving a specific purpose. A luxury-bus is not just a normal bus, it has many other components which change its physiology.There may also be one or more components with better efficiency, like a sports car has a turbo engine, compared to a car that has ordinary engine.

***Manuscript received Dec 5, 2011*.**

***Chandni Singh**, Department of Computer Science, Indian Institute of Information Technology, Design and Manufacturing, ., (e-mail: chandnisi@iiitdmj.ac.in).Jabalpur, India,*

***Rishabh Srivastava**, Department of Computer Science, Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, India, (e-mail: rishabhsr@iiitdmj.ac.in).*

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science & Information Technology (NCETCSIT-2011)

Such variations are primarily because of variations in usage areas i.e. its purpose changes altogether. As in the case of a bus, the purpose of a luxury-bus is not just transportation of an entity from one place to another but laying more emphasis on providing comfort to the entity transported along with its transportation.

- Artifacts may be divided on the basis of certain division criteria e.g. the division criteria medium can be used to differentiate between a car and an airplane. Also car and bus lie under the same category and hence components like wheel are common to all land vehicles. Thus artifacts may be distinguished from one another on the basis of certain criteria so as to allow maximum inheritance of properties and features once a hierarchy is designed. These criteria differ from one domain to another, i.e. we need to define different parameters for maximizing the inheritance in different domains and on different levels.

WordNet discusses a wide variety of artifacts as follows:

- *Facility*
  Something designed and created to serve a particular function and to afford a particular convenience or service
  e.g. catering facilities; toilet facilities; educational facilities
- *Enclosure*
  Artifact consisting of a space that has been enclosed for some purpose
  e.g. open kitchen, playground
- *facility, installation*
  A building or place that provides a particular service or is used for a particular industry
  e.g. the assembly plant is an enormous facility equipped for joining various parts to form a complex artifact.
- *Instrumentality, Instrumentation*
  An artifact (or system of artifacts) that is instrumental in accomplishing some end
  e.g. an ignition key is an instrument for the action 'starting a car'
- *Lemon, Stinker*
  An artifact (especially an automobile) that is defective or unsatisfactory
  e.g. a defective tube light
- *Opening*
  a vacant or unobstructed space that is man-made
  e.g. "they left a small opening for the cat at the bottom of the door"
- *Restoration*
  Some artifact that has been restored or reconstructed e.g. "the restoration looked exactly like the original"
- *Structure, construction*
  A thing constructed; a complex entity constructed of many parts
  e.g. "the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons"
- *Surface*
  The outer boundary of an artifact or a material layer constituting or resembling such a boundary
  e.g. "there is a special cleaner for these surfaces"; "the cloth had a pattern of red dots on a white surface"

IV. Population of Artifacts

A. Manual Population

Manual population of artifacts can be done. Though this method is time taking, it may give high accuracy if the people assigned the job of extracting artifacts from a given list are trained and can clearly differentiate between an artifact and a non-artifact. For this we worked on approximately 7,500,000 articles from the Wikipedia dump. Titles of Wikipedia articles of Wiki data were extracted by a code. This list of titles is then subjected to manual cleaning by trained individuals. The results of a manual method were found to be wrong quite a number of times depending on the accuracy as achieved by an individual. So, the accuracy for the number of correct artifacts differs from person to person. Also this method is laborious and time consuming. Owing to all these reasons, this method was not very efficient.Manual checking by an expert gave 23 correct results for artifacts separated from a list of 1000 titles from Wiki dump. Assuming same distribution throughout the data, we can extract almost 1,72,500 artifacts by this method from Wikipedia alone. This method may be used from lists available on other encyclopedias and corpuses as well.

B. Population using lists available on the internet

A list of artifacts can also be populated from various lists available in different books, vocabulary lists, advertisements, etc. To increase the number of such artifacts, we find places wherein artifacts are highly talked about. Lists on a specialized category may be searched for on the internet. For this first we need to identify the specific domain on which artifacts are to be searched for. Next we look for web pages which may provide lists of artifacts in this area. We may use a web crawler for this so as to improve our results and fasten the process. E.g. extraction of artifacts under the following domains could be done from the given sites.

i. *Medicines*
http://www.medicinenet.com/medications/alpha_a.htm

ii. *Grocery*
http://www.scribd.com/doc/2240296/Ultimate-Grocery-List

iii. *Modes of transport*
http://wiki.answers.com/Q/List_of_all_the_Different_modes_of_transport_from_A-Z

iv. *Building and structure*
http://en.wikipedia.org/wiki/List_of_buildings_and_structures

v. *Clothes*
http://www.manythings.org/vocabulary/lists/a/words.php?f=clothes_1

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science & Information Technology (NCETCSIT-2011)

vi. *Shoes* http://www.enchantedlearning.com/wordlist/shoes.shtml

vii. *Computers* http://readerszone.com/facts/10-different-types-of-computers.html

viii. *Furniture* http://www.enchantedlearning.com/wordlist/furniture.shtml

ix. Cocktails, mock tails and soft drinks

x. Sauce

Other than these we got a large list of artifacts from http://www.wipo.int/classifications/nivilo/pdf/eng/locarno/loc1eng.pdf. This list alone contains about 10000+ artifacts.

The above links together help in accumulating over 23000 artifacts. Thus this may prove to be quite an efficient method provided we can find relevant lists or links for the required kind of artifacts.

C. Population using the artifacts available in Wikipedia

Next we tried to automate the process of finding artifacts from Wikipedia. This can be done by analyzing the information in the first paragraph of an article. We use certain cues (surface text patterns) for discerning the artifacts from the list of titles using information from the article [3]. Various surface text patterns have been formed both for selecting a title as an artifact as well as negating the possibility of a title being an artifact.

(i) Cues which were found to confirm a title as an artifact include: -

a. <artifact> is (a kind of/used as a/ a form of) <artifact>

b. <artifact> is (an item/ an item of/ used as an item of/ made/ manufactured/ prepared/ built/ constructed/ a tool/ a concept/ a complex concept)

c. made <artifact>

d. <artifact> is * tool

(ii) Cues which can be used to reject a title as an artifact include: -

a. <non-artifact> is (a technology/ used technologically/ a unit of/ the measure of/ a belief/a group of/the king of/ a soldier/ a region)

here king may be replaced by (president/ first */ emperor/ king/ queen/ knight/ duke/ duchess/ ruler/ prince/ princess/ writer/ tourist/ person/ soldier/ nurse/ family/ team/ composer/ son/ daughter/ uncle/ aunt/ philosopher/ actor/ director/ newspaperman/ computer scientist/ engineer/ scientist/ bishop/ anthropologist/ linguist/ leader/ founder/ town/ Principality/ publisher/ company/ filmmaker) and other roles which can be played by a human being or other common nouns used for personalities. Also region may be replaced by (technology/ technique/ concept/ unit/ measure/ region/ corporation/ committee/ resource/ set of/ quantity/ game/ sport/ indoor game/ outdoor game/ cocktail/ mock tail/ mission/ model/ field/ list/ film/ movie/ station/ city/ country/ county/ sequence/ faction/ any living thing process/ formula/ equation/ plant/ animal/ human/ species/ cells/ character/ branch) and so on extending the list.

b. <non-artifact> is a * soldier

c. * is a *** composer of

d. history of <non-artifact>

e. <non-artifact> history

f. In <non-artifact>,

(iii) Patterns could be used to definitely reject the titles for artifacts if they contain words like: -

King, Duke, Family, Emperor, Politics, College, University, School, Conference, Committee, Act, Model, Politics, Journal, geography of, battle, war of, book, Principality, Demographics, Pattern, Revolution, etc. More such domains and words can be found by bootstrapping of the process including more and more data.

Observation and Problems:

(i) Patterns for negation are very large in number and many more compared to those for selection. So we see that using negation first helps in rejecting a large list of titles and hence leaving a comparatively shorter list to check for. Now selection can be used, but it helps with only a limited number of titles. Even now a large number of titles from which selection still needs to be done are left.

(ii) Formation of selection as well as rejection surface text patterns for derived noun titles like ‘anarchism’ is very difficult.

(iii) Computer related artifacts are difficult to differentiate. e.g. “They made a program” is a sentence for which .if we take a code as a program, then we will not take it as computer related things are intangible. Therefore, even though the word ‘made’ specifies that a particular word, ‘program’ in this case, is an artifact but it does not confirm it being physical or abstract.

(iv) There are certain Wikipedia articles having just one sentence in the first paragraph. This is not enough for its identification as an artifact. e.g.

a. Demographics of Hong Kong
b. Dalhousie University
c. Head end
d. Epigram
e. College Football

It is difficult to extract information regarding these.

(v) Efficient ranking of STPs can be done using techniques described in [2].

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science & Information Technology (NCETCSIT-2011)

D. Population of the artifacts available in the dictionary

We downloaded the gcide xml dictionary. GCIDE is the GNU version of the Collaborative International Dictionary of English. This was further used to populate the artifact list after identifying basis for selection or rejection of words based on the field it belongs to and the meaning of the words.

The format for the word 'abalienate' is given as:

```
<p><hw>Ab*al"ien*ate</hw><pr>(&abreve;b*
&amacr;l"y&eitalic_;n*&amacr;t;      94,
106)</pr>,    <pos>v.    t.</pos><ety>[L.
<ets>abalienatus</ets>,    p.    p.    of
<ets>abalienare</ets>;  <ets>ab</ets>  +
<ets>alienus</ets> foreign, alien. See
<er>Alien</er>.]</ety><sn>1.</sn><fld>(C
ivil Law)</fld><def>To transfer the title
of    from    one    to    another;    to
alienate.</def><br/>
```

The above format means that the word Abalienate is a verb, has other forms, abalienatus and abalienare, which mean the same as this. The word is then shown to be broken into its root i.e. ab+alienus and then the meaning of alienusis given as foreign. Then a reference to consult alien is also given. One word may have more than one meaning. The information is followed by the '<sn>1.</sn>', this means that its the first meaning. Then field of this word is given as civil law, specified within <fld> tag. This information is then followed by the definition of abalienate i.e. to transfer the title from one to another, which has been specified within the <def> tag.

Parameters Used:

- Only nouns have been considered as artifacts as artifacts can only be noun and no other part of speech.
- The words which had the field as physics, zoology or botany, etc. are rejected. E.g. Shrubs, tiger.
- The words containing keywords like instrument, device and tool are accepted. E.g. Microscope.
- Words which had field as mechanical or medicine, etc were especially kept as they had a high probability of being an instrument or an artifact. E.g. Disprin, stethoscope.
- The words whose meanings initiated with "one", "a person", etc. were eliminated. E.g. Pessimist: - a person who always looks at the negative side of things.
- The words whose meanings initiated with "a state", "a phenomenon", "a unit of", "he", etc. were also rejected as they cannot be an artifact.

Total no of words in the dictionary = 231028
Total no of nouns in the dictionary = 80848
Total no of words extracted = 54241

This method cannot be used for completely creating the required list. After the above step, we need to use any of the other methods for further segregation of the list of artifacts.

One of the major problems we have faced during this entire extraction is time complexity.

(i) Regex matching takes 15 minutes to crawl over a file of 2.5 lakh sentences.
(ii) An artifact match for 5 artifacts takes at least 1 second to compare with 27000 artifacts.

REFERENCES

[1] B. H. Kwasnik, E. D. Liddy, S. H. Myaeng, "Automatic Knowledge Extraction from Dictionary Text: Project Development,"

[2] B. Liu, L. Chiticariu, V. Chu, H.V. Jagadish, F. R. Reiss, "Automatic Rule Refinement for Information Extraction," The 36th International Conference on Very Large Data Bases, September 1317, 2010, Singapore. *Proceedings of the VLDB Endowment,* Vol. 3, No. 1.

[3] D. Ravichandran and E. Hovy, "Learning Surface Text Patterns for a Question Answering System," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 41-47.

[4] E. Oren et al., "Unsupervised named-entity extraction from the Web: An experimental study," Artificial Intelligence, Vol. 165, No. 1. (June 2005), pp. 91-134.

[5] G. Gijs and K. Jan, "Learning Effective Surface Text Patterns for Information Extraction," In 11th Conference of the European Chapter of the Association for Computational Linguistics Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006).

[6] H. Martin, "Automatic Acquisition of Hyponyms from large text corpora," Proceedings of the Fourteenth International Conference on Computational Linguistics.

[7] K. P. Mayee, R. Sangal and S. Paul, "Extraction of Purpose Data using Surface Text Patterns," International Conference on Natural Language Processing and Knowledge Engineering 2010.