Heritage Portals and Heritage Mining:
Synergizing Data Mining and Heritage Preservation under Uncertainty Constraints

Authors:

Milutinovic, V., Salom, J.,Mihajlovic, A., Jelisavcic, V., Ognjanovic, Z., Markovic, Z., Furht, B., Maurer, H.

Abstract:

With ever-increasing demands for collecting, digitalization, and presentation of cultural heritage, a new type of knowledge portals known as Heritage Portals have emerged. The most important features that distinguish such portals from other online knowledge sources are presented. As a consequence, a new type of data mining can be recognized, based on heritage portals as the enabler technology. Key elements of this data mining field are presented, along with two simple and illustrative examples.

1. Introduction

Heritage portals [Korica2012] are specific in many aspects. They include books, documents, historical, and geographical information of encyclopedic character, but also music, still images, and moving images. All this represents a BigData collection with lots of hidden knowledge. Mining the history helps predict future. However, mining from heritage portals is extremely difficult for a number of reasons:

(a) It is multi-dimensional (text, music, images),
and therefore complex.

(b) This complexity is made even higher,
because of the need for different dimensions to synergize.

(c) In addition to the complexity, one has to cope with uncertainty constraints,
arising from the fact that data could have been corrupted, intentionally or unintentionally.

We will review the types of data corruption after we review the basic characteristics of heritage portals, because the heritage portals represent the enabler technology for implementation of heritage mining (heritage mining is best done via cross-corelation of data from heritage portals of different nations). Consequently, characteristics of heritage portals directly influence the types of data corruption.

2. Heritage Portals

Heritage portals must poses specific characteristics. These can be divided into two major categories:

(a) Primary characteristics,

(b) Secondary characteristics.

Primary characteristics of heritage portals are:

- Content is controlled by national institutions established by local governments, with an obligation to worry about digital preservation of national heritage. This is different from the Wikipedia case, where the content is provided by individuals and is controlled by Wikipedia.
- Content is protected by a plethora of legal regulatories, not only by Creative Commons (CC), like with Wikipedia. Each particular document can be protected using a different type of legal regulatory. For example, one document can be copyright-free, while another one can be copyright-protected.
- Content must include titles with short commentaries translated into foreign languages using the culture-oriented translation approach. This means that translations into foreign languages must include hypertext, which makes it semantically understandable to an average reader of the targeted language who grew up in his/her native culture. For example, if a sentence is translated from Japanese to English, and the term “shogun” is used, a hypertext link must be included. Something like that does not exist in Wikipedia, and Google Translate will not produce anything like that.
- Contents must be presented in a way that favorizes quality (and ranking), not quantity (and chaos). For example, the top N artefacts from each contributing institution must be ranked, which enables a viewer in hurry to see quickly the top N contribution of each contributing institution. Something like this does not exist in Wikipedia.

Primary characteristics will always represent differences between Wikipedia on one side and heritage portals on the other side.

One heritage portal that possesses all the above characteristics is Serbia Forum (www.serbia-forum.org). It originated from Austria Forum (www.austria-forum.org). Another member of the family is the heritage portal of Academia Europaea (www.ae-info.org).

Secondary characteristics are also important. They refer to functionalities that enable a more comfortable viewing of the heritage portal contents.Secondary characteristics are not unique to heritage portals. If they exist in heritage portals and not in Wikipedia, two future scenarios are possible:

(a) If they are innovative and good, Wikipedia will soon acquire them, and differences will disappear,

(b) If they are innovativebut bad, Wikipedia will not acquire them, which means that the differences will remain to exist, but will be irrelevant.

Talking about Austria Forum and Serbia Forum, important secondary characteristics include:

- Information must be semantically searchable, by metadata:
  We should be able to search for someone whose name we do not know,
  but we know he was born in Vienna (Wien),was working in physics (Physik),
  and died in Italy (Italien). The search should find Ludwig Boltzman.

- Author must be visible: The source of the text must be always known.
  Either the name of the author including CV,
  or the book/archive it comes from, should be visible.
  The same visibility-related data must be kept on all updates.
- Evolution must be tracked: All versions of the archived item must be reachable,
  from the first update to the last (the present state).
  Note that updating could be dangerous,
  if it destroys previous information and hence history.
- Existence of original commodities: Books are not just a source of information.
  Web books behave like printed books, but also offer some new functions:
  links, bookmarks, … Books should be able to turn into social networks,
  by enabling discussions, comments, ...

Lacking some of the above mentioned characteristics is bringing the decline to Wikipedia [Simonite2013].

These characteristics will exist, also, in a number of other national heritage portals currently under construction. There is an ongoing Europe-wide development effort.

3. Components of the Synergy

Types of mining of interest for heritage portals are listed below. Each type has to be applied independently of others and after that, data from different national heritage portals have to be compared for consistency. The methods include:

a) DataMining (DM) from general databases (DB),
b) TopicMining (TM) from specific databases,
c) ExpertMining (EM) from the Internet-based Social Networks (SN),
d) PsychologyMining (PM) form the Internet-based SNs
   (psychological profiles, mindsets, users' sensibilities, …).
e) DM from still images
   (both indexed and non-indexed, which implies utilization of image understanding tools),
f) DM from moving images (both indexed and non-indexed),
g) DM from Wireless Sensor Networks (WSNs),
   which is of interest for verification of geo-historical data, and
h) DM from the Internet of Things (IoT),
   which is important for observing the present environment.

All the above methods are elaborated in [Milutinovic2013]. The uncertainties that the methods may encounter are:

a) Information is missing (for example an event happened and was not recoded, or was recorded and, later, removed). This type of problem is healed using an imputation method [Mihajlovic2013].
b) Information exists but should not be there (for example, an event had never happened but was recoded). This type of problem is healed using an amputation method.
c) Information was forged (for example an event has happened but the recoding was false). This type of problem is healed using a mutation method.

Organized argumented discussions over the Web using the Argument Web platform and tools [Bex2013] or similar, might be of great help in all these three processes (imputation, amputation, and mutation).

Information from various heritage portals should be compared for consistency. If an inconsistency is noted, cross-correlation, combined with imputation, amputation, and mutation, may lead to a plausible conclusion [Moskowitz2006]. Consequently, a pseudo code of the generalized heritage mining reads as follows:

1. Collect knowledge in a raw form (fill the heritage portal with text, video, audio data,…), determine the data clusters, and form the average distances between different clusters (the stride of the algorithm).
2. Select a point of interest (time point, geo location, person/object of interest…), and declare it the starting point of the extrapolation process;
   check the authenticity of the data involved, using heritage portals of other nations.
3. Initiate the parameters:
   maximum search depth (max_depth),
   current search depth (curr_depth), etc…
   This sets the foundation for comparison of data in different portals.
4. Search all related data-points in the current heritage portal(using standard data mining methods);
   a. If no data was found, or too few data points were found, then:
      i. If curr_depth<max_depth: Find all related heritage portals
         Repeat the step 4 for all related heritage portals
      ii. Go to 5.
   b. For each data-point found, mine the related data-points from related heritage portals.
5. Collect all the found data-points in step 4 along with their originating heritage portal's ids.
6. If data is found missing in the starting heritage portal, do imputation using collected data;
7. If a non-related data point is found in the starting heritage portal, do amputation;
8. If inconsistency is found between the collected data-points, do mutation;
   mutation is best done by combining the data from various heritage portals,
   and by finding "truth" somewhere in the middle area between data of various portals.
9. Extrapolate to a future point and derive a prediction.
10. Check the likelihood level of the prediction, be extrapolating into the past;
    if the extrapolation into the past leads to finding an event that supports the prediction,
    the likelihood of the prediction is considered high.

Generalized algorithms are best understood using an example, so two examples follow.

4. Examples of Heritage Mining

What follows are two examples of heritage mining related to Serbian history. In both cases, the sequence of presented events is obtained using either imputations or amputations or mutations, or a combination thereof.

4.1. The Secret of Numbers 27 and 54

In this example, the heritage mining algorithm implies the following steps, as indicated above:

a. The average distance between generations of data sets in a heritage portal, using a specific number system has to be determined. For example, for the last two centuries, the average distance between generations of people in Serbian families was about 27 years.
b. An important event that drastically changes the entropy of the system has to be found. For example, one can say that the World War I in Serbia started in 1912 with the first Balkan war. Therefore, for this example, the year 1912 can be determined to be the beginning of the coordinating system.
c. The average distance between two data generations (27 in the above example) should be added, iteratively, from the start of the coordinating system to the present time. At each point of iteration, if an important event is missing, imputation candone. Where appropriate, amputation is done. If an event exists that is not on the same level of importance as expected, mutation is done. By applying these three rules, starting from the year 1912, the following sequence is obtained for the case of the recent Serbian history:
    - 1912: The first Balkan war started, which, together with the second Balkan war, represents the Serbian overture to World War I.
    - 1939 = 1912 + 27: The start of the World War II.
    - 1966 = 1939 + 27: The start of the student uprisings, which culminated in 1968.
    - 1993 = 1966 + 27: The start of the culmination of the war in Bosnia.
d. Prediction of the next major political instability in or around Serbia:
    - Around 2020 = 1993 + 27: A major political instability is predicted.
    - Around 2047 = 2020 + 27: Anothermajor political instability is predicted.
e. The quality of the above predictions can be determined by going backwards from the beginning of the coordinated system:
    - 1885 = 1912 - 27: The war between Serbia and Bulgaria.
    - 1858 = 1885 - 27: The war between Russia and Turkey, war with a strong Serbian involvement.
    - 1831 = 1858 - 27: The midpoint of Hatt-i Sharif[1]#1 (1830) and Hatt-i Sharif #2 (1832); both of them had the major impact on the change of the face of Serbia.

[1]The imperial edict in the19th century Turkish empire.

- 1804 = 1831 - 27: The first Serbian uprising against Ottoman empire - the major event in the national history of the last two centuries.

f. The multiples of the number 27 should also be examined:
   - 54 = 2 * 27. Interestingly, the average age of the leaders of each above mentioned political instability was about 54.
   - 81 = 3 * 27. Interestingly, the age of the person who created the major wisdom of the nation, at the time of wisdom creation, was 81.
   - 108 = 4 * 27. Interestingly, the age of the oldest living Serbian, at the time of writing of this paper is 108.

The explanation of the described phenomena is as follows: In some time period, at some geographical area, some people acquire an energy which results in essential advancements (financial strength, population count, etc.). It is natural that these people would like that the political formalisms (division of revenues, border lines, etc...) follow the essential changes. Of course, those on the essential downhills do not want the changes in the political formalism. Consequently, political instabilities happen unavoidably. It is logical that the above described political instabilities happen, with a smaller or a larger amplitude, once per generation; that is why it is important that the average distance between data generations is established in the first step of the heritage mining algorithm presented above.

4.2. The Secret of Numbers 1, 5, 6, 7, and 11

The same algorithm as above could had been used to predict the fall-down of ex-Yugoslavia. By collecting, selecting, initiating, and searching one finds the following (the first 4 steps of the algorithm): Prior to fall-down, national teams of Yugoslavia took part at world cups in five different ball game sports. The teams scored as follows: Position #1 in basketball (5 players on the team), position #2 in volleyball (6 players), position #3 in waterpolo and handball (7 players), position “minus infinity” in soccer - they did not even qualify for the world cup (11 players). Note: More players in the team – worst results of the team.

By checking and extrapolating into the future (steps 5-9 of the algorithm), one can conclude that the team of about 24,000,000 players (the population of ex-Yugoslavia prior to the fall-down) would be blasted away. The conclusion is made even stronger by extrapolating into the past (Monika Seles) and into the far-beyond future (Novak Djokovic): The fact is that the tennis players Monica Seles (one person team in a ball-game, women competition), and the tennis player Novak Djokovic (one person team in a ball-game, men competition), were undisputed #1 for years (this concludes the step #10, which is to determine the likelihood level of the decision).

The above reasoning is based on the fact that getting together into a sport team, and getting together into a statal organization, are correlated.

5. Conclusion

Following are the main viewpoints presented through our work:

Firstly, we present heritage portals as main enablers of heritage mining. By combining data sources from multiple heritage portals, detection and mining of heritage items of interest could be made feasible.

Secondly, we underlined that algorithms for imputation, amputation, and mutation are crucial for mining through heritage portals. Their essence is in cross-corelation of data between various heritage portals, which helps both the starting-point portal and the related portals.

Thirdly, two examples are given to shed light on the issues. It is understood, in both examples, for each major event, heritage portals of related nations also have to be checked, to determine if the events really happened, and if they happened exactly as described in the starting-point portal.

6. References

[Bex2013] Bex, F., et al., “Implementing the Argument Web,” Communication of the ACM, October 2013, Vol. 56, No. 10, pp. 66-73.

[Korica2012] Korica-Pehserl, P., Maurer, H., "Semi-Automatic Information Retrieval and Consolidation with a Sample Application, " Proceeding of the IEEE International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, October 1-8, 2012, pp. 1-6.

[Mihajlovic2013] Mihajlovic, A., "Machine Learning-Based Imputation of Missing SNP Genotypes in SNP Genotype Arrays,"Computational Medicine in Data Mining and Modeling, Springer, New York, 2013, pp.193-231.

[Milutinovic2013] Milutinovic, V., Jelisavcic, V., Mihajlovic, A., et al, “Data Mining from Social and Knowledge Networks, ” Preconference Tutorial, Symposium on Applied Computing, ACM SAC 2013, Coimbra, Portugal, March 2013, pp. 1-60.

[Moskowitz2006] Moskowitz, H., Gofman, A., Beckley, J., Ashman, H., "Founding a New Science: Mind Genomics," Journal of Sensory Studies, Vol. 21, No. 3, 2006, pp. 266-307.

[Simonite2013] Simonite, T., “The Decline of Wikipedia,” MIT Technology Review, Nov./Dec. 2013, Vol. 116, No. 6, pp. 50-56.