Big Tourism Data Analytic Enthalpy in Saudi Arabia

Nasser Nammas Albogami[1], Khalid Allehaibi[1], and Arif Bramantoro[2]
nalbugami@kau.edu.sa kallehaibi@kau.edu.sa asoegihad@kau.edu.sa
[1]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia

Summary

Water- flow like algorithm, CVRP, Great Deluge, Hybrid metaheuristic Saudi Arabia has been a country of superior religious tourism in many decades due to the existence of multi-religious sites, especially Islamic holy sites. Big data has become a key component in Saudi Arabia Tourism industry. Thus, a big data based comprehensive approach is critical to overcome the issues of tourism metrics. In this paper, big data enthalpy, a novel approach in atomizing big data from tourism industry, is proposed to handle big data through a single metric which enables the smooth analysis amongst tourism stakeholders in Saudi Arabia through several experimentation trials. Big data enthalpy is broken down into procedure concentration, analysis softening, operation enhancement, data escalation, data dedifferentiation, and information exposure. Finally, the approach is semantically and syntactically assessed to meet several established aims in Saudi Arabia tourism industry.

Key words:
Big data analytic, Data Enthalpy, Tourism Industry

1. Introduction

Religious sites in Saudi Arabia have been the source of attraction to large scale tourism industry. Research in Saudi Arabia tourism industry relies profoundly on the existence of big data and the aggregation of multiple analyses that exploits them. In other words, obtaining worthful information and even knowledge in the era of the explosion of tourism raw data is the critical element of tourism. Moreover, tourism industry ordinarily harmonizes the different notion of analysis and data that leads to a more sophisticated business process and information.

This paper aims to ensure the collaboration amongst any interested parties, especially during the analysis of tourism phenomenon. However, the task of analyzing tourism activities and data is so arduous that most users are required to be as good as an expert data analyst. In addition, there is not enough longanimity in expecting the last outcome produced by the long process of analysis. Although the last outcome of analysis is obtained, it is too sophisticated to compare with results from other analysis processes. It contains too much information and knowledge. It is also important to note that few options of dashboard are available for tourism stakeholders to access the unfamiliar and intriguing information. Based on informal observation of tourism industry in Saudi Arabia in the last few years, there are several universal characteristics in analyzing tourism industry that are tackled in this paper as follows.

Analysis degree There exists an analysis tool that delivers very hard but trivial processing. Tourism analysis includes a subordinate connection, which is costly to achieve more than two degrees. Soft processing occasionally satisfies users.

Number of procedures A particular analysis tool may consist of several procedures in total to accomplish its objective. In tourism industry, for example, provider analysis involves sentiment analysis, although a frontend user hardly grasps its importance.

Operational quality Tourism analyses generally require a similar amount of operational period for different request, such as working on an uncomplicated question "What are the impacts of the hike of oil price on Saudi tourism in 2019?". This is because tourism analysis consumes the whole data from multiple dimensions, such as time and place.

Data explosion A particular analysis is able to process an explosion of data while another analysis is unable to process the rescaling data. Trust analysis in Saudi tourism industry collects hundreds million crawled websites, although most users utilize provider analysis that require much fewer websites.

Data structuredness Tourism analysis usually defines a specific structure of data input and output. One analysis may provide a plain structure for its output, while another analysis defines a more sophisticated input structure. To aggregate these two analyses is not seamless. However, the legacy tourism industry has no mapping mechanism between analyses.

Intriguing and unfamiliar information The best analysis equipped with big data is not always fruitful in exposing particular information that is intriguing and not yet familiar to tourism industry. If tourism user receives overall analyses and data, there is an opportunity that she misses the intriguing and unfamiliar piece of information. In addition, a little portion of intriguing and unfamiliar information are more useful than the big one.

Experimentation trials The legacy business process for big data analyses aggregation has an impediment to run when dealing with a quality constraint and at the same time within acceptable period. Experimentation enables users to try and modify the analysis in-between business process

Manuscript received Nobember 5, 2019
Manuscript revised Nobember 20, 2019

execution. It means that the final point of execution is unnecessary anymore. It also enables users to choose which data are suitable to accomplish her goal during several trials.

In the next section, the related works will be discussed including the trend in data analytic. Then the proposed tourism platform is presented followed by its enthalpy measurement. A sample interface with its scenario is provided before the experimental evaluation. The last section would be conclusion and future work.

2. Related Work

Tourism industry is currently one of the hottest are for information analysts, such as in [1],[2],[3]; which exploits several well-established guidelines proposed in [4],[5],[6],[7]; categorized as the new wave of knowledge discovery guidelines. However, these works are typically specialized to scientific business process, a legacy business process used for scientific analyses. Time out usually occurs during transfer of big data between analyses. Tourism analyses are different from scientific analyses. They require the involvement of big data transfer between analyses. This is because there are too many unstructured information involved during the tourism process, such as multimedia.

Big data disaggregation is conceptually similar to the ones proposed in [8],[9],[10]. Compared to big data disaggregation, big data enthalpy has different granularities of data. While only the lowest level in big data disaggregation is executable, all levels in big data enthalpy are executable. This approach atomizes big data based on functionality, operation, data escalation, analysis depth, data heterogeneousness, and intriguing information, while big data disaggregation decomposes an analysis based on its functionality only. Big data disaggregation aims at reducing the size of the analysis. On the other hand, big data analytic enthalpy aims at being flexible to any size of the analysis and data, including the increasing number of analyses and data during the execution.

Another related work to big data analytic enthalpy is analysis measurement. Application cohesion is proposed in [11] as a relatedness measurement of application interface. However, it only measures application interface instead of big data involved in application. Moreover, application measurement does not discuss the tendency of changing application to another application, as it normally occurs in tourism industry.

The big data analytic enthalpy is different from entropy of information theory, although many authors tried to exploit the analogy between the expressions of entropy in thermodynamics and those in Shannon's information theory in [12]. The big data analytic enthalpy is not a measure of disorder state of a system. It is proposed as a measure of a change from one tourism analysis to another. Although enthalpy in thermodynamics is strongly related to entropy, big data entropy is not coined in this paper due to the difficulty to find a disorder state in big data analytic, which frequently exists in most tourism industry with its big data analysis.

Experimentation analysis is not a new approach in the information analysis. The authors in [13] proposes this approach to enable scientist to analyze information within small interval and proceed to the next interval in order to support large scale volume of solar image data. The authors in [14] proposes different approach of experimentation in analyzing information. They visualize historical user analysis activities to enable user to revise the query iteratively. The use of web browsing logs and genome scientific data proves the idea of "previous trial result might be better than present one." These two works are accommodated in this paper to support big data and enhance them with tourism analyses to enable the dynamic collaboration between big data analytic during the experimentation activity.

3. Tourism Industry Platform

A new platform is proposed to support various tourism big data analytic that deals with big data assets such as web archive of billion pages, social media and scientific data citation.

As illustrated in Fig.1, the platform accommodates any applications related to the tourism industry in Saudi Arabia. To the best of our knowledge, most applications in tourism industry are based on four main principles, which are generally identified Saudi Arabia. The first principle is that any applications should advance the visualization of tourism activities in Saudi Arabia. Since this activity regularly requires hyper-realistic interaction between users and any application providers, the interaction data are centralized in a big data platform as a single truth.

The second principle is that the tourism application is on top of a dictionary selection which requires multilingual text translation between non-Arabic speaking tourists and Arabic-speaking residents. There is also a chance to involve the interaction amongst non-Arabic speaking tourists, since the holy sites in Saudi Arabia is the most internationally-wide place for pilgrims. The dictionary selection mainly includes the choice of common vocabularies and expressions used by tourists.

The third principle is that the communication tool for tourist should compromise both speech recognition and synthesis to accommodate two-way communication.

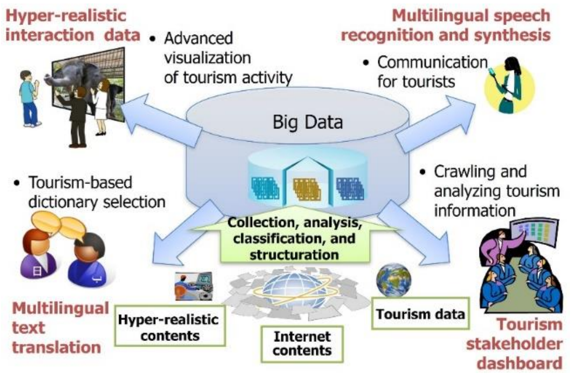

Hyper-realistic interaction data
• Advanced visualization of tourism activity
Multilingual speech recognition and synthesis
• Communication for tourists
Big Data
• Tourism-based dictionary selection
• Crawling and analyzing tourism information
Collection, analysis, classification, and structuration
Multilingual text translation
Hyper-realistic contents
Internet contents
Tourism data
Tourism stakeholder dashboard

Fig. 1 Big Tourism Data Analytic Platform

Finally, the fourth principle is that the tourists should attain an ability to crawl deeply the information surrounding the holy sites and knowledge. Once it is crawled, there is a necessity of analyzing according to the requirement and aim of the tourists. This research focuses on this activity to provide tourism dashboard to the stakeholders.

There are more capabilities of big data analytic accommodated in the platform. Hence, it can handle not only input, output, precondition and effects of big data analytic; but also the quality of data. However, to date, the quality of data processing technique is too network-centric and not flexible for user in particular big data analytic domains. Tourism big data analytic, for example, requires not only operational quality in executing the analysis but also the quality of the analysis results. In addition, this big data analytic requires data intensively in performing the analysis with considerably good results.

The main big data analytic in the platform is tourism analysis. It is considered as a business process that extract valuable information from big data by utilizing text mining, inferencing, semantic reasoning, multimedia information processing, and most importantly, natural language processing, that can be utilized to support tourism campaign. Some researchers use the term knowledge discovery on database or data mining, however, it is preferred to use the term tourism analysis to emphasize valuable information extraction and big data on tourism industry. In tourism, user usually starts from a hypothesis or a question which is considered as a goal by this platform. Eventually, users update their goals or hypotheses after acquiring new information during analysis.

Existing tourism tools and application are synchronized in the platform to improve their operation and scalability. Improving operation is by reducing the end-to-end analysis time and improving scalability is by scaling vertically and horizontally. Scaling vertically is by enabling big data analysis and scaling horizontally is by allowing heterogeneous data analysis. The platform is required to accommodate both scalabilities, although these two terminologies are not necessarily used throughout this paper.

Currently, the major information analysis tools incorporated for tourism in Saudi Arabia are WISDOM in [15], Torishiki in [16], and Ikkyu in [17]. WISDOM is a knowledge mining search engine equipped with trust analysis as well as sentiment, trouble, provider, and appearance analysis. Torishiki is an information connection analysis tool for conceptual terms available on web that currently covers more than two millions words. Ikkyu is an open domain question answering system from 600 million web pages equipped with voice.

Everything as a big data analytic is the solution for connecting these tools. Once those tools are considered as a business process in big data analytic, it is possible to compose those processes to fulfil new requirements from user to analyze tourism-related information. However, composing those processes is not straightforward. They have different output standards. WISDOM analysis, for example, has a complex information as an output of user's statement. Ikkyu analysis has a completely different format in its output, it retrieves a list of related concepts. Torishiki analysis has an analysis collection of trouble, method, tools, and so forth. This motivates us to have a new approach in dealing with not only big data but also fine-grained analysis.

4. Big Data Analytic Enthalpy

Big data analytic enthalpy is an approach to atomize a heavy tool or big data analytic application into smaller forms of APIs that supports bigger data and at the same time satisfies what user requires most. Big data analytic enthalpy wraps the APIs as one atomic analysis and compose with other analyses which requires bigger data to collaborate with based on a measurable quality. This paper extends the previous work of service atomization presented in [18] and service enthalpy presented in [19] into big data analytic in tourism industry.

Big data analytic enthalpy also atomizes data for each data store. An archive application is utilized to store and partition large scale data to enable further processing by tourism industry. Archive application then indexes the data. In this platform, archive application is built in the distributed system for some domains, such as crawled web pages, social network messages, and geo-sensing data.

In general, big data analytic enthalpy is a measure of the total efforts required to separate all atomic data from existing applications or tools in tourism industry. The same notion as in thermodynamics is borrowed to measure enthalpy, big data analytic enthalpy cannot be measured directly. The change of big data analytic enthalpy, ΔH, is the one that can be measured as defined in Eq. 1.

$$\Delta H = H_{\text{final}} - H_{\text{initial}} \qquad (1)$$

where ΔH is the enthalpy change, H_{final} is the enthalpy once the system has been atomized into smaller data, and H_{initial} is the enthalpy of the data before being atomized.

Thus, the change of big data analytic enthalpy, ΔH, is more useful to run several tourism analyses since it provides direct calculation between analyses. Throughout this paper, however, big data analytic enthalpy is mentioned to refer the enthalpy change. In general, big data analytic enthalpy is formally defined in Eq. 2.

$$H = U + aT \tag{2}$$

where H is big data analytic enthalpy, U is the enthalpy of original system before being atomized, a is the data required by the system and T is the processing time to run the system. If the product for differentiation is employed, Eq. 3 is obtained.

$$\begin{aligned} \mathrm{d}H &= \mathrm{d}(U + aT) \\ &= \mathrm{d}U + \mathrm{d}(aT) \\ &= \mathrm{d}U + (a\mathrm{d}T + T\mathrm{d}a) \end{aligned} \tag{3}$$

The equation shows that the process of atomization on a system needs a change of the amount of data required by the system and processing time to run the system. If the amount of data is raised in a constant time, there is a need of additional effort as the system changes. However, if the amount of data is raised and the required processing time is equally conserved, there is no additional effort needed to change the original system.

By using big data analytic enthalpy, it is possible to increase the quality of analysis in step by step manner, since particular data can be dynamically substituted in a business process during execution. It is possible in achieving the same level of quality of data as provided by the tool and application before atomized, by composing the atomic data in one business process until it reaches the same level of functionality. In this paper, there are six enthalpies proposed for tourism industry.

4.1 Enthalpy of Procedure Concentration

Enthalpy of procedure concentration is a number of procedures provided by an atomic big data analytic that can be eliminated or added to transform into another simpler or more complex atomic data. If the number of procedures reduces, its enthalpy change ΔH_{pc} is negative, and if the number of procedure increases, ΔH_{pc} is positive.

Initially, there is a personal consultation to tourism analysts to arbitrarily determine the enthalpy of procedure concentration and calculate it by using Eq. 2. Similar procedures are clustered into the same library and quantified in the different enthalpy as listed on Table 1. Due to the page limitation, only several procedures that utilized in the scenario are discussed in the detail.

Table 1: Big Data Analytic and Their Enthalpies ΔH_{pc}

Big data analytic procedure	Library	ΔH_{pc}
Lemmatization	Preprocessing	1
C4.5	Classification	2
SimpleKMeans	Clustering	2
J48	Classification	3
RandomForest	Classification	4
ADTree	Classification	4
Classifier	Classification	4
Concept Dictionary	Association	4
Attribute Analysis	Association	4
Method Analysis	Association	5
Subordinate Connection	Association	6
Provider Analysis	Classification	7
Link Analysis	Classification	7
Trouble Analysis	Association	7
Sentiment Analysis	Classification	8
Trust Analysis	Classification	10
WISDOM	Association	15
Torishiki	Association	15
Ikkyu	Association	15
Stream Concordance	Visualization	20
MathPlotter	Visualization	20

4.2 Enthalpy of Analysis Softening

Enthalpy of analysis softening is an analysis depth of atomic big data analytic that can be cut or added to transform big data analytic into another atomic big data analytic with softer or harder analysis. If a new big data analytic has softer analysis than the original big data analytic does, its enthalpy change ΔH_{as} is negative, and if a new big data analytic is harder, ΔH_{as} is positive.

Basically, hard analysis applies as much tourism knowledge as possible to analyze user's query. However, the definition of analysis depth might be different from one big data analytic to another. Big data analytic of trouble, method, and question-answer have the amount of resulted knowledge as a parameter to determine the enthalpy of analysis softening. The analyses of trust, provider and appearance have two modes of knowledge delivered to users, i.e. summary and detail information. Therefore, two different enthalpy values are required for these analyses. Enthalpy of analysis softening is supported by the platform by implementing rule-based platform. Typically, hard analysis is rule-based. It uses rules to store and describe knowledge as much as possible to analyze information. Other analyses, such as sentiment analysis, require a softer analysis that can be delivered by using enthalpy of analysis softening.

4.3 Enthalpy of Operation Enhancement

Enthalpy of operation enhancement is the response time difference between one new atomic big data analytic and another analytic before being transformed. When the response time reduces, its enthalpy change ΔH_{oe} is negative. On the other hand, when the additional response time is required, ΔH_{oe} is positive.

Enthalpy of operation enhancement uses one metric of quality of data, i.e. data processing time. The historical processing time of each data is recorded and averaged. The historical average of processing time is used as enthalpy change of operation enhancement to create a new atomic data. Enthalpy of operation enhancement is useful to avoid *time out* that usually occurs during transfer of big data between analyses. It is also useful as a prediction for user to avoid some analyses that are not in a good operation.

4.4 Enthalpy of Data Escalation

Enthalpy of data escalation is a size adjustment of the required data size when creating a new atomic big data analytic transformed from an existing one. If the data size reduces, its enthalpy change ΔH_{de} is negative, and if the data size increases, ΔH_{de} is positive.

The term of data escalation is coined for this enthalpy, because the larger data is not necessarily the higher recall and precision of analysis they have. In trouble analysis, for example, there are four enthalpy values of data escalation, based on databases consisting of 10,000, 22,000, 30,000 and 300,000 trouble nouns. Current database version contains 22,000 trouble nouns which has more than 5 million associations. The more accurate the tourism analysis tool has, the higher chance it lowers the operation. For example, the under-developing tourism analysis tool of 100 million associations might be slower for tourism analysis with the current one with 5 million associations.

Historical database is preserved in order to avoid an important information lost after human annotator performs a data cleaning. Several big data analytics are built based on enthalpy change of data escalation. Each big data analytic handles different database. Trouble analysis, accordingly, is atomized into four big data analytics with ΔH_{de} of 12,000, 8,000, and 220,000 between analyses. *WISDOM* analysis has a different way. It is atomized into some analyses to deal with 1 billion pages of crawled data pool, others are with 200 million pages of text data pool, and the rest are with 120 million pages of analysis target pool. By using enthalpy of data escalation, the different size of big data can be efficiently analyzed by big data analytic in this platform.

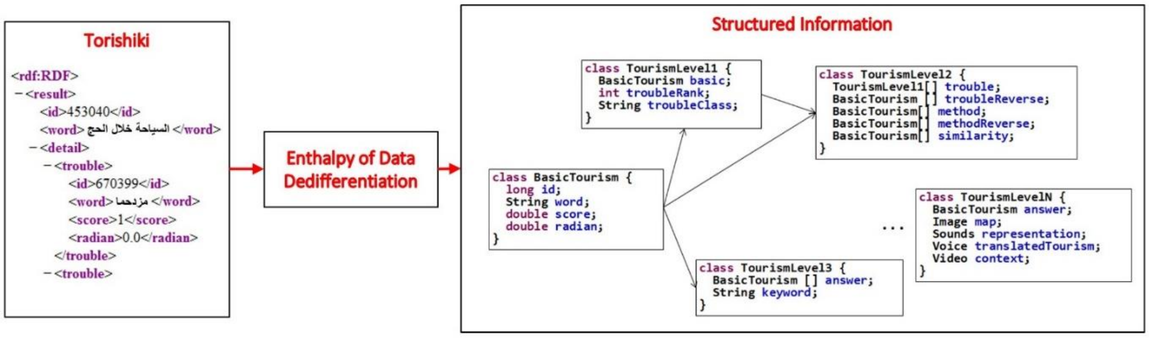
4.5 Enthalpy of Data Dedifferentiation

Enthalpy of data dedifferentiation is for data structure of an atomic big data analytic that is dedifferentiated in order to be used by another atomic data. If a new big data analytic is more structured in its data than the original big data analytic is, its enthalpy change ΔH_{dd} is negative, and if a new big data analytic has less structured data, ΔH_{dd} is positive.

To support heterogeneousness and unstructured data, it starts from the most structured data and reduce its enthalpy of data dedifferentiation so that the new data can be handled by different big data analytic. The most structured data structure is useful to share basic information amongst big data analytics. In tourism industry, for example, there is a common data structure amongst big data analytic enabling easier aggregation of several analyses by sharing common information. As illustrated in Fig. 2, all big data analytic tools have basic tourism analysis result in a list of *id*, *related word*, *score* used to group related words based on their association score; and *radian* used to measure the distance between the query and its related word. Hence, any output formats, including semantic format, are able to be structured by utilizing enthalpy of data dedifferentiation.

Trouble analysis uses *TourismLevel*1 data structure to add more information of seriousness (*troubleRank*) and to group it based on a verb that represents worse scenario (*troubleClass*). Another big data analytic which requires five kinds of analyses, trouble, trouble reverse, method, method reverse, and similarity; uses *TourismLevel*2 to retain those information.

Question-answering *Ikkyu* analysis uses *TourismLevel*3 to add keywords from user's question. Other levels of data structure can be added based on existing data structures to support fine-grained tourism. In the future, it can accommodate more unstructured information, such as images of context around the tourism topics, tourism sounds and videos.



```
<rdf:RDF>
 – <result>
     <id>453040</id>
     <word> السياحة خلال الحج </word>
   – <detail>
     – <trouble>
         <id>670399</id>
         <word> مزدحما </word>
         <score>1</score>
         <radian>0.0</radian>
       </trouble>
     – <trouble>
```

```
class TourismLevel1 {
  BasicTourism basic;
  int troubleRank;
  String troubleClass;
}
```

```
class TourismLevel2 {
  TourismLevel1[] trouble;
  BasicTourism [] troubleReverse;
  BasicTourism[] method;
  BasicTourism[] methodReverse;
  BasicTourism[] similarity;
}
```

```
class BasicTourism {
  long id;
  String word;
  double score;
  double radian;
}
```

```
class TourismLevelN {
  BasicTourism answer;
  Image map;
  Sounds representation;
  Voice translatedTourism;
  Video context;
}
```

```
class TourismLevel3 {
  BasicTourism [] answer;
  String keyword;
}
```

Fig. 2 Data structure dedifferentiation for each big tourism data analytic

Trouble analysis uses *TourismLevel*1 data structure to add more information of seriousness (*troubleRank*) and to group it based on a verb that represents worse scenario (*troubleClass*). Another big data analytic which requires five kinds of analyses, trouble, trouble reverse, method, method reverse, and similarity; uses *TourismLevel*2 to retain those information.

Question-answering *Ikkyu* analysis uses *TourismLevel*3 to add keywords from user's question. Other levels of data structure can be added based on existing data structures to support fine-grained tourism. In the future, it can accommodate more unstructured information, such as images of context around the tourism topics, tourism sounds and videos.

4.6 Enthalpy of Information Exposure

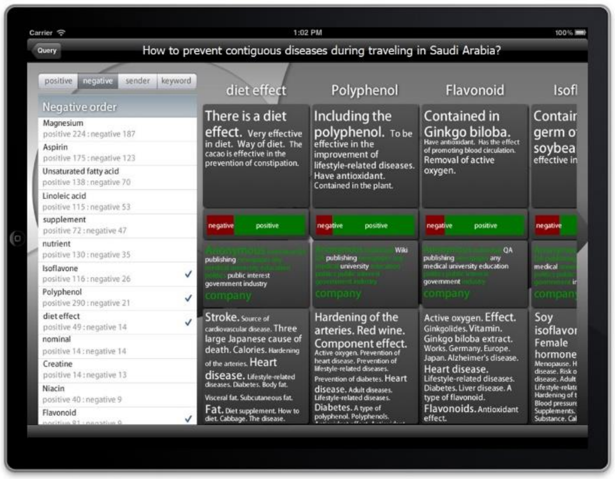
Enthalpy of information exposure is the different possibility of exposing a useful but unfamiliar tourism analysis results from one atomic big data analytic to another big data analytic. If the possibility increases, its enthalpy change ΔH_{ie} is positive, and if not, ΔH_{ie} is negative.

The experimental results in [15],[16],[17] are collaborated to calculate enthalpy of information exposure. The experiment in [15] revealed 70% of useful but unfamiliar information in trust analysis and 80% of those in sentiment analysis. The experiment in [16] revealed 31.1% of useful but unfamiliar information in trouble analysis, 20% of those in method analysis, and 6.7% of those in analysis tools. The experiment conducted in [17] has no information regarding to the exposure for question answering analysis, however, the same enthalpy is performed for trouble analysis since these two are almost similar. To get a standard enthalpy of information exposure, each experimental results are normalized.

These six enthalpies are used in the platform to create new analyses from existing tourism industry. More specifically, big data analytic enthalpy is used to write rules for aggregating several analyses. For example, the rule that aggregates several analyses from the lowest enthalpy to the highest user is outlined during the request. Another rule is when big data analytic is too slow, it should be replaced with another big data analytic that has a lower enthalpy. By defining the enthalpy based rules, user has a recommendation to refine his goals or hypotheses during aggregation and execution of tourism data analytic.

Another rule is when a change of analysis is multiplied by some factors, enthalpy must be multiplied by the same factors. For example, a trouble analysis runs three times before combined with trust analysis. In this case, an enthalpy of trouble analysis must be multiplied by three. The last rule is when a change of big data analytic is reversed, the sign of enthalpy must also be reversed. For example, a method analysis that is changed into a reversed method analysis has enthalpy calculated as the same amount of original enthalpy but with a negative value.

Two tools from Weka4WS in [5] and Orange4WS in [4] are collaborated into the platform in two fashions. First, those tools are required to enhance the analysis result from internal tools available in the platform. For example, the list of related concepts delivered by trouble analysis is very long that needs to be sorted and cut. It is not possible to apply general sorting algorithm on the association score between related concept and query since it has also radian score to express the semantic relatedness between one related concept and others. Therefore, a robust clustering algorithm is provided by an external analysis to obtain the concepts with the highest association score for each clustered concepts.



How to prevent contiguous diseases during traveling in Saudi Arabia?

Negative order

Magnesium positive 224 : negative 187	
Aspirin positive 175 : negative 123	
Unsaturated fatty acid positive 138 : negative 70	
Linoleic acid positive 115 : negative 53	
supplement positive 72 : negative 47	
nutrient positive 130 : negative 35	
Isoflavone positive 116 : negative 26	✓
Polyphenol positive 290 : negative 21	✓
diet effect positive 49 : negative 14	✓
nominal positive 14 : negative 14	
Creatine positive 14 : negative 13	
Niacin positive 40 : negative 9	
Flavonoid positive 81 : negative 9	✓

diet effect

There is a diet effect. Very effective in diet. Way of diet. The cacao is effective in the prevention of constipation.

Stroke. Source of cardiovascular disease. Three large Japanese cause of death. Calories. Hardening of the arteries. Heart disease. Lifestyle-related diseases. Diabetes. Body fat. Visceral fat. Subcutaneous fat. Fat. Diet supplement. How to diet. Cabbage. The disease.

Polyphenol

Including the polyphenol. To be effective in the improvement of lifestyle-related diseases. Have antioxidant. Contained in the plant.

Hardening of the arteries. Red wine. Component effect. Active oxygen. Prevention of heart disease. Prevention of lifestyle-related diseases. Prevention of diabetes. Heart disease. Adult diseases. Lifestyle-related diseases. Diabetes. A type of polyphenol. Polyphenols.

Flavonoid

Contained in Ginkgo biloba. Have antioxidant. Has the effect of promoting blood circulation. Removal of active oxygen.

Active oxygen. Effect. Ginkgolides. Vitamin. Ginkgo biloba extract. Works. Germany. Europe. Japan. Alzheimer's disease. Heart disease. Lifestyle-related diseases. Diabetes. Liver disease. A type of flavonoid. Flavonoids. Antioxidant effect.

Isofl

Contair germ o soybea effective in

Soy isoflavor Female hormone Menopause. H disease. Risk o disease. Adult Lifestyle-relat Hardening of t Blood pressure Supplements. Substance. Cal

Fig. 3 A scenario of tourism by adopting experimentation flow

Second, external analysis is collaborated to replace internal one when it is not available or underperformed. For example, sort-and-cut graph of analysis result can be very slow since it involves many analyses. This slow big data analytic is substituted with external big data analytic, such as *PathFinder* in [20], a graph simplification to prune a weighted graph. All enthalpies of external big data analytic are measured to streamline the integration with internal analysis.

5. Experimentation Flow

It is not possible to try particular big data analytic in a regular business process and return to big data analytic later when it completes or returns no error. Current big data analytic aggregation technology has a limitation that requires the whole big data analytic running in one business process without leaving any possibilities to examine the temporary analysis results in the middle of business process execution. Moreover, a tourism analysis usually requires big data that is not possible to be interrupted in the middle of the execution.

The users of tourism industry usually face the challenge of finding useful information from large scale data. They have to continue analyzing the data until they get the required information. They have to pick up some parts of the data, analyze them, and integrate the result with previous ones in a particular order. A complete paradigm, such as CRISP-DM in [21], needs additional steps of cleaning, constructing and reformatting the data. CRISP-DM also emphasizes the need of considering the aim as the main part in the analysis process [22]. In most cases, however, user's aim is still ambiguous. The user might not have the target keywords at the beginning. The user has to try the analysis results and data to have more obvious aim and keywords.

Fig. 3 shows an interface in mobile device that allows experimentation in analyzing tourism. In experimentation flow, the user can change big data analytic dynamically in the middle of big data analytic business process execution without waiting for the completion of the execution. Experimentation is a suitable approach to analyze data stream such as messages from social network system. Big data analytic enthalpy is initially proposed to support experimentation flow by providing big data analytic at different levels of granularity and providing a measurable recommendation during the execution of experimentation flow.

In experimentation flow, the computational burden of analysis on the shoulder of the client is shifted away. Most analysis processes are stored in server side and most functionalities are provided by the platform. Its objective is to enable user to try all possibilities of analysis and return to the previous intermediate result by utilizing experimentation flow. To test the usability of big data analytic enthalpy approach, a tourism application interface is required to support experimentation flow in any devices, such as mobile device.

Note that there is no legacy business process in this scenario that requires a control flow between big data analytics. In this interface, the big data analytic is automatically executed by firing rules in the tourism platform. This scenario dictates a rule of starting big data analytic from the lowest enthalpy to the highest one in order to guide user during the decision making on which big data analytic that might be useful for the current trial. In another scenario, there are several rules utilizing big data analytic enthalpy, such as recommending the data in interval of big data analytic enthalpy based on historical enthalpy used by the same user or group of users, and refactoring big data analytic enthalpy for a big data analytic that is executed repetitively.

In this example, user can input his query to the experimentation interface. The application searches data and analyses based on user query and shows them based on big data analytic enthalpy. From the recommended data and analyses, user chose social media data to analyze. The platform anticipates action by recalculating all user choice and existing big data analytic enthalpy. Therefore, archive for social media is recommended as well as other as well as other analyses with higher big data analytic enthalpy. When a user puts big data analytic in a higher position of flow area (*Ranker*), it executes big data analytic in parallel with another big data analytic in a lower position (*Array parser*) and combine the results to big data analytic.

The platform supports a detection of unreliable analysis. *Torishiki*, for example, is very slow in processing a list of cutoff social media messages. A rule is dictated to suggest another analysis with similar enthalpy of data escalation but with a better enthalpy of operation enhancement (i.e. *Subordinate connection analysis*) when the processing time of *Torishiki* is beyond the limit.

In experimentation flow, the session mechanism is required to enable user to store her temporary analysis results in case she wants to return to the previous state of analysis. The session is also required to enable user to refine her query string iteratively. By using this session, all user experiences in finding information in the experimentation approach are stored in the platform. Another advantage is that the best experience of a user in finding a useful information can be shared with other users, although we plan to address the evaluation of the sharing result as a future work.

6. Evaluation and Discussion

To simplify the evaluation of big data analytic enthalpy, six enthalpies are reorganized in two ways, semantic and syntactic disparity. Semantic disparity is used to measure how different knowledge is acquired between analyses, while syntactic disparity is used to measure statistical natural language processing. Enthalpies of data escalation, information exposure, and analysis softening are considered as semantic disparity, whereas enthalpies of procedure concentration, data dedifferentiation and operation enhancement are considered as syntactic disparity. In this experimentation evaluation, the disparity of semantic and syntactic can be reduced to a minimum value close to zero which is considered as the accomplishment of user's aim in analyzing tourism.

There are four aims used for evaluating big data analytic enthalpy: formulating a tourism trend, analyzing user query regarding to tourism, proving a tourism hypothesis and brainstorming new ideas in evaluating tourism. These aims are based on the perspective of the stakeholders of Saudi Arabia tourism industry during the decision making. All big data analytic enthalpies are capitalized in the tourism analytics and data to meet each aim. The result is shown in Fig. 4 (note that not all analyses and data are shown).

Each aim starts from analyses and data with the most suitable enthalpy and collaborates with other analyses and data. This process repeats until it reaches the aim which is indicated by a minimum disparity value in semantic and syntactic. For example, the first aim (formulating a trend) runs s_1 which is *stream concordance* to visualize the trend of user keyword in social network messages. It then runs s_2 which is *trouble analysis* to find other related concepts in trouble domain. The user chooses some related concepts and some messages to be classified by s_3, *SimpleKMeans*, to cluster the information as a trend.

Someone may argue that several trivial activities may arise due to experimentation. For example, sorting a correlation score of related concepts in *Ikkyu* analysis tool may eventually be executed, although it has been already sorted. However, more experimentation can be conducted to find out the related concepts are already sorted. This is because that they look like not being sorted in the first trials.

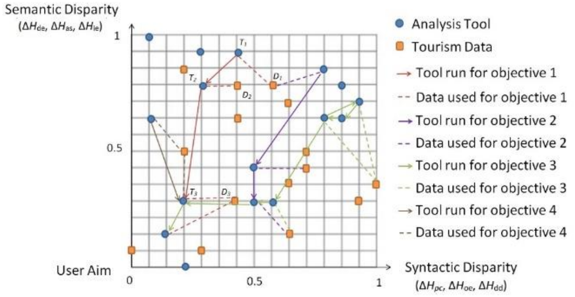


Fig. 4 Evaluation of six enthalpies

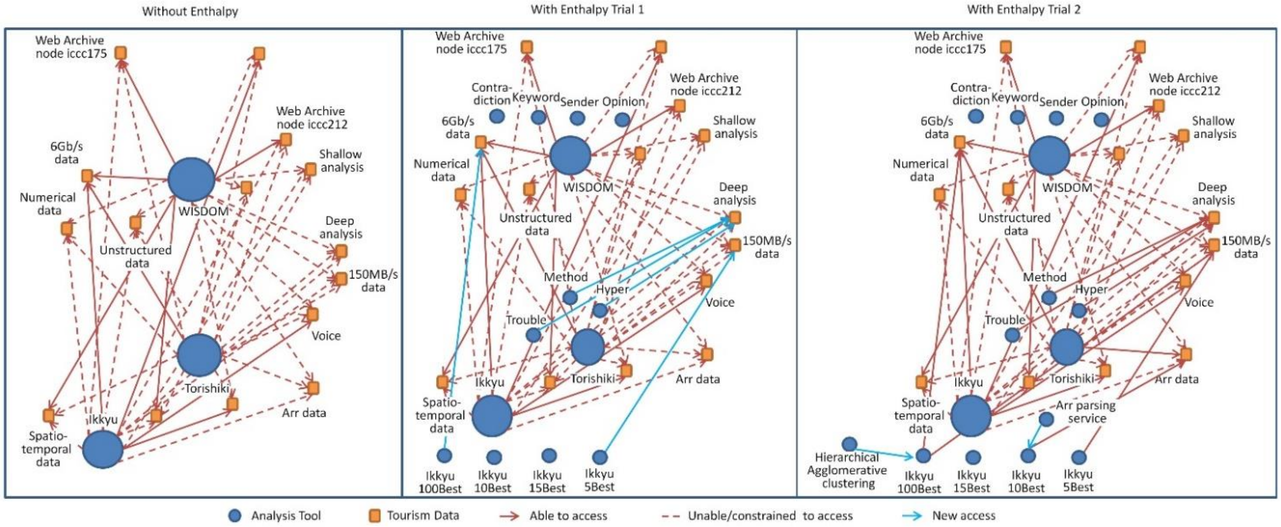


Fig. 5 Big data enthalpy trials

Another advantage is that the new access may get published by running big data analytic enthalpy for tourism industry in several trials. Fig. 5 illustrates three trials of running several tourism analytic tools with tourism data which leads to two possibilities, the access is enabled or not. By running the big tourism data enthalpy in the first trial, the new accesses are created to adjust the need of existing data, such as Ikkyu 100Best, Ikkyu 5Best, Method, Trouble, Hyper analysis tools. It is important to note that these tools are not verified as a final results, hence, the possibility of false positive remains is well recognized.

The slicing and drilling analysis tools for dense and coarse data is different from the ones for less dense and coarse data. In the second trial of big tourism data enthalpy, the hierarchical agglomeration and classification is found to be very useful for more specific questions, such as into what, where, and why into. Hence, these two tools are accommodated in the trial. However, the end-to-end analysis time is not evaluated in detail in this paper. It is assumed that the overall analysis time with big data analytic enthalpy is better than the one with bigger tools as also identified in [23].

7. Conclusion

A novel advancement in tourism industry was presented to enable better handling of big data. The tourism analysis is atomized into many other analyses at finer granularity based on big data analytic enthalpies, a measure of effort in creating an atomic analysis from huge application and data. Six enthalpies are used: enthalpy of procedure concentration, analysis softening, operation enhancement, data escalation, data dedifferentiation, and information exposure. This approach of big data analytic enthalpy is assessed through an experimental user interface to examine the usability of the platform in Saudi Arabia. The result demonstrates that there are several trials conducted in step by step manner that dynamically enabled different access to new analysis tools based on the enthalpy of the tourism data. It is argued that big data analytic enthalpy is applicable not only in tourism domain, but also in other big data analytic domains in the future.

References

[1] V. Shapoval, M. C. Wang, T. Hara, and H. Shioya, “Data mining in tourism data analysis: Inbound visitors to japan,” Journal of Travel Research, vol. 57, no. 3, pp. 310–323, 2018.

[2] G. Kurian and H. Chi, “Predict florida tourism trend via using data mining techniques,” in Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact. ACM, 2017, p. 68.

[3] K. Pitchayadejanant and P. Nakpathom, “Data mining approach for arranging and clustering the agro-tourism activities in orchard,” Kasetsart Journal of Social Sciences, vol. 39, no. 3, pp. 407–413, 2018.

[4] V. Podpecan, M. Zemenova, and N. Lavrac, “Orange4WS environment for Service-Oriented data mining,” The Computer Journal, vol. 55, no. 1, pp. 82-98, 2012.

[5] D. Talia, P. Trunfio, and O. Verta, “The Weka4WS framework for distributed data mining in service-oriented grids,” Concurrency and Computation: Practice and Experience, vol. 20, no. 16, pp. 1933–1951, 2008.

[6] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure, “myExperiment: a repository and social network for the sharing of bioinformatics workflows,” Nucleic Acids Research, vol. 38, pp. 677–682, 2010.

[7] I. Taylor, M. Shields, I. Wang, and A. Harrison, “The triana workflow environment: Architecture and applications,” Workflows for e-Science, pp. 320–339, 2007.

[8] H.-K. Yang and H.-S. Yong, “Distributed parafac decomposition method based on in-memory big data system,” in International Conference on Database Systems for Advanced Applications. Springer, 2019, pp. 292–295.

[9] X.-Y. Liu and X. Wang, “Ls-decomposition for robust recovery of sensory big data,” IEEE Transactions on Big Data, vol. 4, no. 4, pp. 542–555, 2017.

[10] I. Notarnicola, R. Carli, and G. Notarstefano, “Distributed partitioned big-data optimization via asynchronous dual decomposition,” IEEE Transactions on Control of Network Systems, vol. 5, no. 4, pp. 1910– 1919, 2017.

[11] M. Perepletchikov, C. Ryan, and Z. Tari, “The impact of service cohesion on the analyzability of Service-Oriented software,” IEEE Transactions on Services Computing, vol. 3, no. 2, pp. 89–103, 2010.

[12] C. Shannon and W. Weaver, The mathematical theory of communication. University of Illinois Press, Urbana, Illinois, USA, 1949.

[13] E. Stolte and G. Alonso, “Approximated trial and error analysis in scientific databases,” Information Systems, vol. 28, no. 1-2, pp. 137– 157, 2003.

[14] K. Nishimura and M. Hirose, “The study of past working history visualization for supporting trial and error approach in data mining,” Human Interface & the Management of Information, pp. 327–334, 2007.

[15] T. Kawada, S. Akamine, D. Kawahara, Y. Kato, Y. I. Leon-Suematsu, K. Inui, S. Kurohashi, and Y. Kidawara, “Web information analysis for open-domain decision support: system design and user evaluation,” in Joint Workshop on Web Quality, 2011, pp. 13–18.

[16] K. Torisawa, S. De Saeger, J. Kazama, A. Sumida, D. Noguchi, Y. Kakizawa, M. Murata, K. Kuroda, and I. Yamada, “Organizing the web’s information explosion to discover unknown unknowns,” New Generation Computing, vol. 28, no. 3, pp. 217–236, 2010.

[17] I. Varga, K. Ohtake, K. Torisawa, S. De Saeger, T. Misu, S. Matsuda, and J. Kazama, “Similarity based language model construction for voice activated Open-Domain question answering,” in International Joint Conference on Natural Language Processing, 2011, pp. 536–544.

[18] A. Bramantoro, T. Kamada, M. Tanaka, Y. Murakami, and K. Zettsu, “Towards service atomization for analyzing information,” in 2012 IEEE 19th International Conference on Web Services (ICWS), 2012, pp. 676– 677.

[19] A. Bramantoro, “Service enthalpy for analyzing cybercrime,” in Anti-Cybercrime (ICACC), 2015 First International Conference on. IEEE, 2015, pp. 1–6.

[20] A. Vavpetic, V. Batagelj, and V. Podpecan, “An implementation of the pathfinder algorithm for sparse networks and its application on text networks,” in International Multiconference Information Society, 2009, pp. 236–239.

[21] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide. Chicago: SPSS Inc., 2000.

[22] G. Piatetsky-Shapiro, “Crisp-dm, still the top methodology for analytics, data mining, or data science projects,” KDD News, 2014.

[23] H., Wolfram, M. Fuchs, and M. Lexhagen. "Big Data Analytics for Tourism Destinations," Encyclopedia of Information Science and Technology, Fourth Edition, pp. 349-363, IGI Global, 2018.