# การวิเคราะห์บทความอัตโนมัติ โดยใช้ กระบวนการภาษาธรรมชาติ

## Literature Analyzed Using Natural Language Processing

**กรมวุฒิ นงนุช[1*] อนุชา ซาเฮาะ[1] และ สุวุฒิ ตุ้มทอง[1]**

Krommavut Nongnuch[1*], Anucha Zahao[1] and Suwut Tumthong[1]

**บทคัดย่อ**

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบช่วยสรุปบทความวิจัยที่มีการตีพิมพ์แบบอัตโนมัติ โดยทำการดึงเอาบทความจากแหล่งเผยแพร่บทความที่มีการตีพิมพ์แบบสาธารณะเช่น Science Direct, PubMed, IEEE เป็นต้น ใช้วิธีทาง XML, Call API, Web Service, Web Content เป็นต้น ใช้การดึงบทความแบบอัตโนมัติ เป็นประจำทุกวัน เพราะบทความมีการตีพิมพ์เผยแพร่ทุกวันทำให้ได้ความรู้ที่ใหม่อยู่เสมอจากการเขียนโปรแกรมแบบ Crontab Daily ระบบมีการทำงาน 3 ขั้นตอนหลัก อันดับแรก, จะรวบรวมบทความวิเคราะห์ RSS ฟีดสำหรับการเผยแพร่บทความใหม่จาก PubMed, ScienceDirect และ Sprinkerlink อันดับที่สองการวิเคราะห์บทความจะใช้ขั้นตอนวิธีการรับรู้การแยก RSS แท็กและข้อมูลที่เกี่ยวข้องในเนื้อหา สกัดและเพิ่มข้อมูลลงในฐานข้อมูลที่สามการวิเคราะห์บทความ ทำการสร้างการแจ้งเตือนสรุปรายวันจากฐานข้อมูล นอกจากนี้ข้อมูลจากการวิเคราะห์บทความ จะสามารถใช้ได้ในเว็บไซต์รวมถึงความสามารถในการค้นหาและรายงานการวิเคราะห์ เว็บไซต์ที่มีรายการสรุปและข้อมูลรวมทั้งการเชื่อมโยงไปยังบทความวารสารด้วย

**คำสำคัญ** : Machine Learning, XML, Call API, Web Service, Web Content, Crontab Daily

**Abstract**

This research aims to develop an automatic literary analysis. Textual summaries from the resource of publications such as Science Direct, PubMed, and IEEE were automatically generated in XML, Call API, Web Service and Web Content format. Due to the research articles were daily published, the new knowledge from scheduler program and Crontab Daily was obtained. System works through 3 main steps. First, Literature Analyzed will collect RSS feeds for new nomenclatural declaration from PubMed, ScienceDirect, and Sprinkerlink. Second, Literature Analyzed will apply Recognition algorithms to parse the RSS to tag and associated information within the content, extract and add the information into its database. Third, Literature Analyzed will generate the summarize alert daily for the subscriber. In addition, the information from Literature Analyzed database will be available at the website including search capabilities and analytical reports. The website contains the list of summarize and its information including the link to the journal article

**Keywords:** Machine Learning, XML, Call API, Web Service, Web Content, Crontab Daily

[1] คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสวรรณภูมิ ศูนย์ นนทบุรี 11000

[1] Faculty of Science and Technology,Rajamangala University of Technology, Suvarnabhumi, Nonthaburi, 11000, Thailand

* Corresponding author. E-mail: krommavut.n@rmutsb.ac.th

## บทนำ

ในต่างประเทศการศึกษางานวิจัยมีความสำคัญในการนำความรู้ไปพัฒนาประเทศทั้งภาคอุตสาหกรรมและภาคการศึกษา ดังนั้นการศึกษางานวิจัยที่มีการเผยแพร่จึงมีความสำคัญต่อการพัฒนาประเทศที่กำลังพัฒนาอย่างยิ่งเพราะ จะทำให้ประเทศที่กำลังพัฒนามีความสามารถเท่าเทียมประเทศที่พัฒนาแล้ว แต่ปัญหาของการศึกษางานวิจัยนั้นต้องมีความเข้าใจและศึกษากระบวนการต่างๆ จากการอ่านงานวิจัย ซึ่งต้องใช้เวลาในการอ่านบทความวิจัยและสรุปประเด็นของผู้เขียนบทความเพื่อเผยแพร่สาธารณะ จึงได้มีการวิจัยเครื่องมือช่วยในการศึกษางานวิจัยใช้เทคนิคทางด้านภาษาธรรมชาติ ฝึกให้คอมพิวเตอร์มีการเรียนรู้ ภาษาของมนุษย์ หรือภาษาเขียน แล้วให้คอมพิวเตอร์ ทำการสรุปเนื้อหาออกมาเฉพาะประเด็นสำคัญ

จากเหตุผลดังกล่าวผู้วิจัย จึงได้เกิดแนวคิดทำเครื่องมือ เพื่อสรุปประเด็นการเผยแพร่บทความสาธารณะจากแหล่งที่เชื่อถือได้เพื่อให้ผู้สนใจบทความลดระยะเวลาในการสรุปประเด็นต่างของบทความโดยใช้เทคนิดกระบวนการภาษาธรรมชาติ(Natural Language Processing) เข้ามาเป็นเครื่องมือช่วยในการสรุปประเด็นต่างๆ ของบทความ

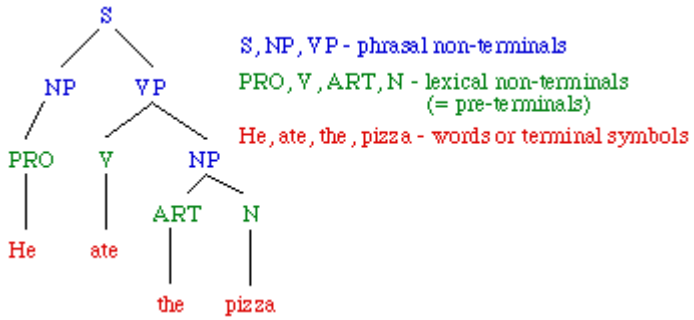
## ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

**ระบบประมวลภาษาธรรมชาติ** (Natural Language Processing System)

การประมวลผลภาษาธรรมชาติเป็นระบบที่จะช่วยทำให้คอมพิวเตอร์เข้าใจ ภาษาธรรมชาติของมนุษย์เพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษาเขียน รูปแบบความรู้ที่คอมพิวเตอร์สามารถนำ ไปใช้งานได้เช่น การสรุปบทความ การตอบโต้ทางแป้นพิมพ์ ระบบตัดคำ เป็นต้น

**นิยามที่ 1การวิเคราะห์ในเชิงโครงสร้าง** (Syntactic Analysis)

**การตรวจสอบโครงสร้างทางไวยากรณ์ ตำแหน่งของคำประเภทต่าง ๆ ที่รวมกันเป็นประโยค เช่น** He ate the pizza




**ภาพที่** 1 วิเคราะห์ในเชิงโครงสร้าง

**นิยามที่ 2 การวิเคราะห์ในเชิงความหมาย** (Semantic Analysis)

เป็นการตรวจสอบความถูกต้องในเชิงความหมาย ของประโยค โดยประโยคที่วางกลุ่มคำชนิดต่างๆ ตามโครงสร้างไวยากรณ์ จะมีความหมาย อย่างใดอย่างหนึ่งแน่นอนแต่ในบางครั้งประโยค ที่กำลังพิจารณาอาจจะเขียน ถูกต้องตามหลักไวยากรณ์ แต่มีความหมายกำกวมหรือเป็นความหมายที่เป็น ไปไม่ได้ หรือไม่ให้ความหมายอะไรเลย[2] “The house eat the boys” จะเห็นว่าประโยคนี้โครงสร้างของประโยคถูกต้องตามหลัก

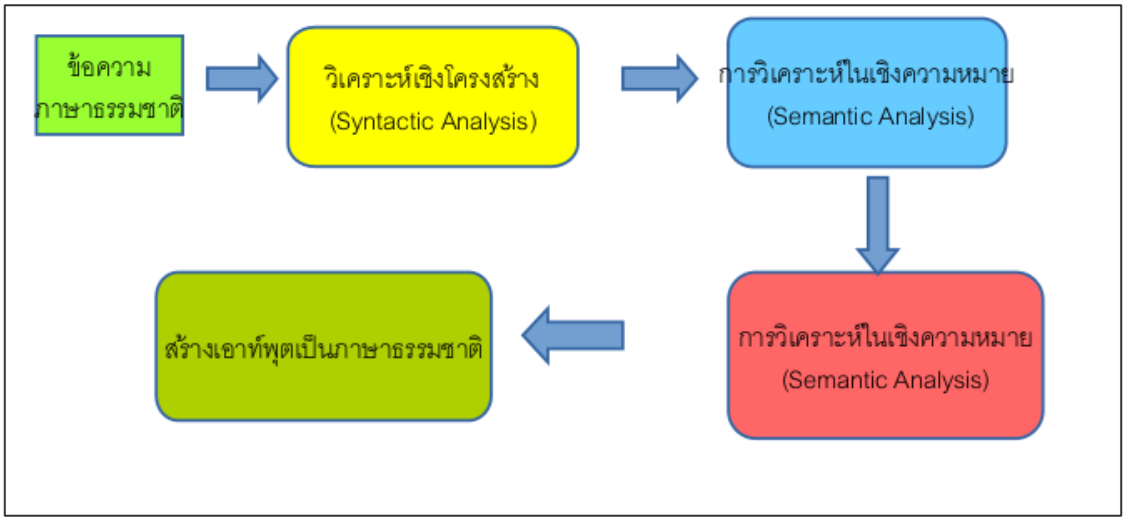
ไวยากรณ์คือแต่เมื่อวิเคราะห์ดูความหมายแล้วเห็นว่าประโยคนี้มีความหมาย ที่เป็นไปไม่ได้ ในเมื่อบ้านเป็นสิ่งไม่มีชีวิตจึงทำกริยา “กิน” ไม่ได้

**ประโยค** = **นามวลี** +**กริยาวลี** **นามวลี** = **คำนำหน้านาม** +**คำนาม**
**กริยาวลี** = **กริยา** + **นามวลี**

**ภาพที่** 2 วิเคราะห์ในเชิงความหมาย

**นิยามที่** 3 **การวิเคราะห์ในเชิงตีความ** (Pragmatic Analysis)

ประโยค ที่เขียนออกมาบางครั้งก็อาจจะไม่ได้มีความหมายตรงตามข้อความเช่น เราอยู่ที่สถานีรถไฟและกำลังกังวลว่าขณะนี้ เวลาเท่าไรรถไฟใกล้จะออกหรือยังแต่เราไม่มีนาฬิกา พอดีหันไปเห็นคนข้าง ๆ กำลังดูตารางเวลาการเดินรถเช่นเดียวกัน เราเลยหันไปถามว่า “Do you have a watch?” ถ้าเราได้คำตอบว่า “yes” หรือ “no” แสดงว่าคำตอบที่ได้ผิด เพราะคำตอบที่เราต้องการจริงๆ คือ เวลา ณ ขณะนี้




**ภาพที่** 3 ประมวลผลภาษาธรรมชาติ

**การรวบรวมบทความสาธารณะ** (Summarize Publication Search)

**นิยามที่** 1เอกซ์เอ็มแอล (XML : Extensible Markup Language)

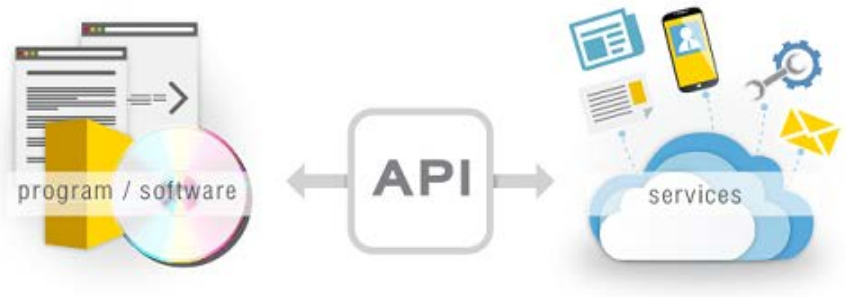
[2]เป็นภาษาที่ใช้เน้น (มาร์กอัป) ส่วนที่เป็นข้อมูล โดยสามารถกำหนดชื่อแท็ก (Element) และชื่อแอตทิบิวต์ ได้ตามความต้องการของผู้สร้างเอกสาร xml โดยเอกสารนั้นจะต้องมีความเป็น Well-formed ส่วน DTD และ Schema จะมีหรือไม่มีก็ได้ ขึ้นอยู่กับว่ามีผู้ใช้เอกสารนั้นมากน้อยแค่ไหน เอกสาร xml จึงเป็นแค่เท็กซ์ไฟล์ชนิดหนึ่ง ที่มีแท็กเปิดและเท็กปิดครอบข้อมูลไว้ตรงกลางเท่านั้น ทำให้เอกสาร xml ถูกใช้ในการติดต่อกับระบบที่ต่างกัน เนื่องจากความง่ายในการสร้างเอกสาร การนำเอกสาร xml ไปใช้งาน จะสนใจแต่ข้อมูลที่ถูกเน้นด้วยแท็กมากกว่า

```
<pubmed>
 <title>Computer Science</name>
 <id>123456789</id>
</pubmed>
```

**ภาพที่** 4 Tag XML

**นิยามที่ 2 เอพีไอ** (API : Application Programming Interface)

[2] ช่องทางการเชื่อมต่อ ,ช่องทางหนึ่งที่จะเชื่อมต่อกับเว็บไซต์ผู้ให้บริการ API จากที่อื่น เป็นตัวกลางที่ทำให้โปรแกรมประยุกต์เชื่อมต่อกับโปรแกรมประยุกต์อื่น หรือเชื่อมการทำงานเข้ากับระบบปฏิบัติการ

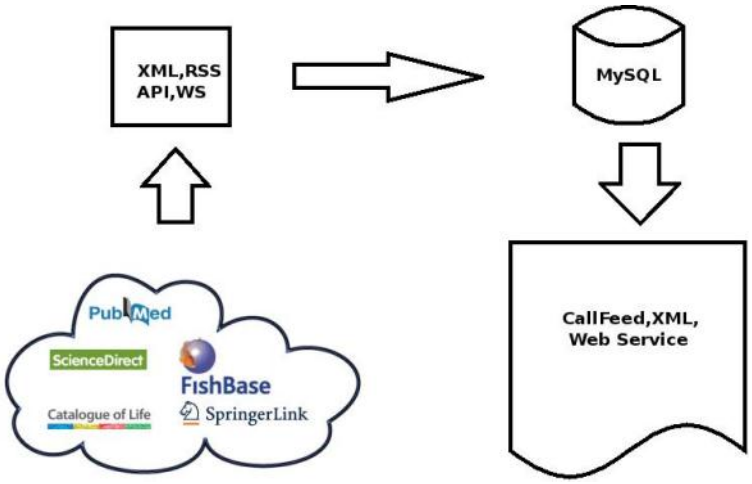



**ภาพที่** 5 การเรียกใช้งาน API

**วิธีการศึกษา** (Methodology)

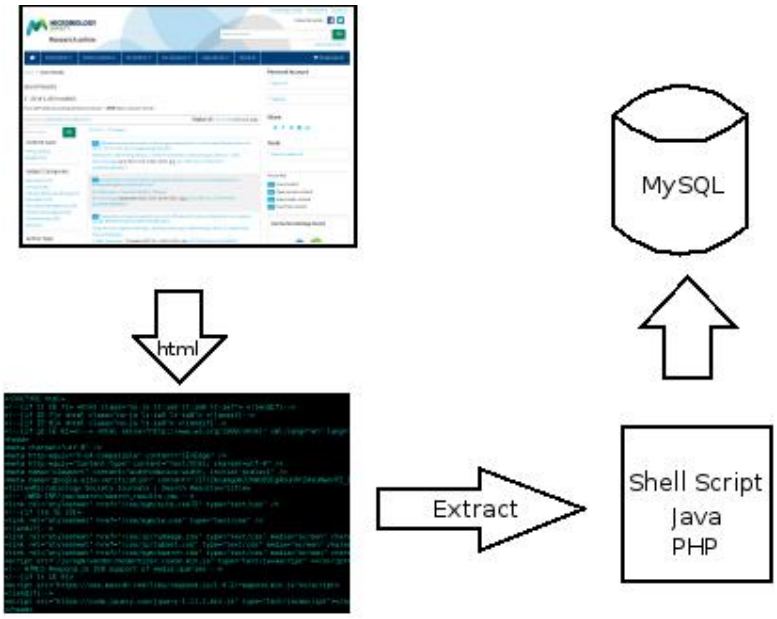
**การทำกระบวนการภาษาธรรมชาตินั้นต้องมีการฝึกให้คอมพิวเตอร์เรียนรู้**

3.1 ดึงข้อมูลอัตโนมัติ การดึงข้อมูลอัตโนมัติจะใช้ XML, API, Web Content ในการดึงข้อมูลเพื่อทำเป็นระบบอัตโนมัติ



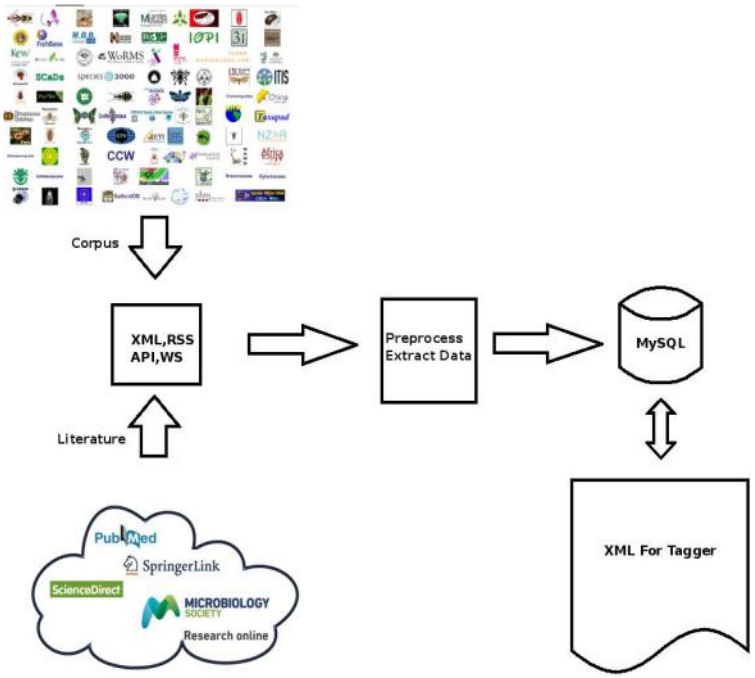

**ภาพที่** 6 การดึงข้อมูลอัตโนมัติ

3.2 การตั้งสคริปเพื่อดึงข้อมูลแบบวนซ้ำ เพื่อนำข้อมูลเข้าระบบทุกวัน(Cron Daily) ใช้ภาษา Shell ,JAVA , PHP ทำงานร่วมกัน ทำให้เกิดข้อมูลที่มีการปรับปรุงตลอดเวลา และเป็นการสะสมข้อมูลเพื่อทำเหมืองข้อมูลอีกด้วย




**ภาพที่** 7 การตั้งสคริปเพื่อดึงข้อมูลแบบวนซ้ำ

3.3 การสร้าง Corpus นำสถิติจากข้อมูล ที่สร้างขึ้นจากแหล่งข้อมูลที่เชื่อถือได้ มาฝึกให้คอมพิวเตอร์สามารถเรียนรู้ข้อมูลและสามารถแยกกลุ่มของข้อมูลได้ โดยใช้เทคนิค C4.5 Algorithm, CRF Algorithm เป็นต้น [1]




**ภาพที่ 8** การสร้าง Corpus

3.4 การวิเคราะห์ในเชิงตีความ (Pragmatic Analysis)การวิเคราะห์ในเชิงตีความจะใช้ Open Source OpenNLP เป็นตัวดำเนินการตรวจสอบโครงสร้างทางไวยากรณ์เกี่ยวกับการวางตำแหน่งของคำนาม กริยา คำบุพบท ฯลฯ ที่รวมเป็นประโยค แยกแยะความถูกต้องทางความหมายของประโยค ประโยคที่ถูกวางแบบตามโครงสร้างไวยากรณ์จะมีความหมายอย่างใดอย่างหนึ่ง แน่นอน แต่บางครั้งความหมายที่ได้เป็นความหมายที่กำกวม หรืออาจไร้ความหมาย

วางแผนทางกระบวนการภาษาธรรมชาติ [2]

3.4.1 ใช้ฐานความรู้เป็นหลักในการประมวลผลภาษา (parser) โดยใช้กฎไวยากรณ์ที่สร้างโดยผู้เชี่ยวชาญ

3.4.2 ใช้สถิติจากข้อมูล ที่สร้างขึ้นจาก ชุดฝึก corpus เทคนิคทางสถิติ และใช้ชุดข้อมูลนั้นฝึกให้คอมพิวเตอร์สามารถคิดเป็น เช่น HMM Algorithm, C4.5 Algorithm, CRF Algorithm etc.

3.4.3 การผสมผสานกันระหว่าง **ใช้ฐานความรู้**และ ใช้สถิติจากข้อมูลเพื่อใช้ฝึกคอมพิวเตอร์ให้รู้จักข้อมูลที่มากขึ้นและวิเคราะห์ข้อมูลได้

## ผลการศึกษาและอภิปรายผล

การวิเคราะห์ผลลัพธ์เมื่อทำการส่งข้อความไปในระบบจะทำการ ตัวสอบไวยากรณ์ที่สำคัญของข้อความนั้นและสกัดเฉพาะวลีที่สมบูรณ์ออกมาจัดอันดับ

Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. This was due to both the steady increase in computational power (see Moore's Law) and the gradual lessening of the dominance of Chomskyan theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing.[3] Some of the earliest-used machine learning algorithms, such as decision trees, produced systems of hard if-then rules similar to existing hand-written rules. However, Part-of-speech tagging introduced the use of Hidden Markov Models to NLP, and increasingly, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to the features making up the input data. The cache language models upon which many speech recognition systems now rely are examples of such statistical models. Such models are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data), and produce more reliable results when integrated into a larger system comprising multiple subtasks.

**ภาพที่** 9 ทดลองส่งบทความเข้าระบบ

**ตารางที่** 1 ผลของกระบวนการภาษาธรรมชาติเพื่อหาลำดับความสำคัญของไวยากรณ์จากข้อความ

| ลำดับ | ไวยากรณ์ |
|---|---|
| 1 | Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. |
| 2 | Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. |
| 3 | [3] Some of the earliest-used machine learning algorithms, such as decision trees, produced systems of hard if-then rules similar to existing hand-written rules. |
| 4 | The cache language models upon which many speech recognition systems now rely are examples of such statistical models. |
| 5 | Such models are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data), and produce more reliable results when integrated into a larger system comprising multiple subtasks. |

จากภาพที่ 9 ทดลองนำข้อมูลเข้าระบบสามารถจำแนกได้ตามตารางที่ 1 โดยมีการจัดอันดับไวยากรณ์ไว้ทำให้ผู้สนใจดูแค่ประเด็นที่สำคัญได้

## สรุป

การใช้กระบวนการภาษาธรรมชาติ ในการสรุปข้อมูลของเอกสารที่สำคัญจะช่วย ลดระยะเวลาในการศึกษา งานวิจัยที่มีความยากในการอ่านหรือแปลความหมายของบทความนั้น การนำกระบวนการภาษาธรรมชาติที่ดีควรมีเป้าหมายที่มีประโยชน์ต่อสาธารณะและช่วยอำนวยความสะดวกในการใช้

## เอกสารอ้างอิง

[1] S. Atdag, and V. Labatut. (2013). A Comparison of Named Entity Recognition Tools Applied to Biographical Texts, (pp. 228-233). 2nd International Conference on Systems and Computer Science.

[2] Y. Fanjiang and Y. Syu, (2014). Semantic-based automatic service composition with functional and non-functional requirements in design time: A genetic algorithm approach, 56(3), pp. 352-373, Information and Software Technology.

[3] J. Sangers, F. Frasincar, F. Hogenboom, and V. Chepegin, (2013).Semantic Web service discovery using natural language processing techniques, 40(11), pp. 4660- 4671, Expert Systems with Applications.

[4] K. Gunaratna, S. Lalithsena, and A.P. Sheth, (2014). Alignment and Dataset Identification of Linked Data in SemanticWeb, 4 (2), pp. 139-151, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

[5] John Cuzzola, Jelena Jovanovic, Ebrahim Bagheri, Dragan Gasevic, (2015). **Evolutionary fine-tuning of automated semantic annotation systems,Volume 42 Issue 20**, Expert Systems with Applications.

[6] Boris A. Galitsky, (2013).Transfer learning of syntactic structures for building taxonomies for search engines, Volume 26, Issue 10, pp 2504–2515,Engineering Applications of Artificial Intelligence.