وفاء بن التركي
أستاذة بالمعهد العالي للترجمة بالجزائر
باحثة في مجال الترجمة الآلية والمعالجة الآلية للغات الطبيعية
رئيسة جمعية الترجمة الآلية لولاية الجزائر
البريد الالكتروني: obenterki@hotmail.com

رقم الهاتف المحمول: 00213770604760

موضوع البحث : المعطيات الضخمة Big Data : تحديات الانفجار المعرفي وعلاقته باللغة

كلمات مفتاحية
المعطيات الضخمة Big Data، الترجمة الآلية Machine Translation، استرجاع المعلومات Information retrieval، تحليل المعطيات الضخمة BIG Data Analysis ،المتون اللغوية Linguistics Corpora ، تقانة اللغة Language technology .

الملخص

لقد أدى الانفجار المعرفي المنقطع النظير الذي ميز هذا القرن إلى تغيير جذري في كل مجالات الحياة، إذ أصبحت هذه الأخيرة مرتبطة ارتباطا وثيقا بالمعلومات. والسبب في ذلك هو وتيرة إنتاج المعارف المتسارعة إلى جانب الأحجام الهائلة للمعلومات المتداولة بِلُغَاتٍ مختلفة عبر الشابكة والتي واكبت حقبة المعلومات.

وتتميز المعطيات الضخمة Big Data المتناثرة هنا وهناك عبر الشابكة ببنية معقدة وغير منتظمة، مما صعب عملية استكشاف المعارف، وبات استرجاعها وتصنيفها وتخزينها يمثل تحديا كبيرا أمام البرمجيات الحديثة. وبما أنه ليس بمقدورنا فصل المعلومات عن سياقها فإنها حتما ترتبط بمنظومة اللغة التي تنتمي إليها.
لقد خصصت الجزء الأول من البحث لتعريف المعطيات الضخمة والبرمجيات المكرسة لتحليل هذه المعطيات أما الجزء الثاني فقد تم تكريسه لعلاقة المعطيات الضخمة بتقانة اللغة Language technology.

وقد تم تكريس الجزء الثالث لدور الترجمة الآلية في استيعاب المعطيات الضخمة من خلال إدماجها في عملية تحليل المعلومات على اختلاف أنواعها. وينتهي البحث بالخاتمة التي تتضمن بعض التوصيات.

1- تعريف المعطيات (البيانات) الضخمة Big Data

تعددت التعاريف التي أسندت المعطيات الضخمة، ومن ضمنها :

2-1 تعريف جارتنرز مارفن أدريان [1] Gartner's Marven Adrian

« ليس في وسع الأدوات البرمجية أو الحواسيب الرائجة الاستعمال توفير بيئة مناسبة لحصر وإدارة ومعالجة المعطيات الضخمة ومن ثم وضعها في متناول المستخدمين في غضون مدة زمنية مقبولة».

3-1 تعريف معهد ماكانزي[2] العالمي للمعطيات الضخمة

« يشير مصطلح المعطيات الضخمة Big Data إلى مجموعات من المعطيات يفوق حجمها قدرات برمجيات قواعد البيانات النظامية على الحصر والتخزين
والإدارة والتحليل ».

إن الضخامة الأُسية Exponential التي تميز هذا النوع من المعطيات لا تقتصر على حجمها الكبير فحسب بل يتعدى ذلك إلا سمات أخرى :

1− **الحجم Volume**

تتميز **المعطيات الضخمة Big Data** بأحجام هائلة من المعلومات تتضاعف أحجامها على الشابكة بوتيرة متسارعة. وقد كانت الهيئات والمؤسسات والشركات فيما مضى أهم مصدر للمعلومات إلا أن الانفجار المعرفي الذي ميز العقد الأخير غير المفاهيم السابقة وأوجد مصادرا جديدة للمعلومات على غرار الأفراد والآلات الذكية.

فبعد أن كانت إدارة الشركة أو المؤسسة هي المصدر الرئيسي للمعلومات المتداولة داخل الشركة (أو المؤسسة أو الهيئة) أصبح للشركاء والعملاء (الزبائن) نصيبهم من المعلومات من خلال استحداث طرق جديدة للإدارة تستوجب اشراك العملاء التعرف عل توجهاتهم ومواكبتها من خلال استطلاعات الرأي Surveys أو غيرها من البرمجيات الحديثة.

ومن جهة أخرى، تم استحداث برمجيات ذكية يتم تفعيلها بصفة آلية تقوم بتزويد الشبكة الداخلية (الموظفين) أو الخارجية (المتعاملين) بمعلومات محددة على غرار الرسائل القصيرة أو البريد الإلكتروني الآلي.

ومثال ذلك مشروع الرصد الإجمالي للمعلومات الفلكية[3] والذي يستعمل جهاز تيليسكوب ضخم يقوم برصد الظواهر الفلكية في السماء من خلال المسح الآلي وتسجيل هذه الظواهر مدة عشرية كاملة ، وسيتطلب الامر استرجاع صور وتسجيلات فيديو بحجم يصل إلى

---

[1] **Gartner's Marven Adrian: Q1, 2011 Teradata - Magazine**

[2] McKinsey Global Institute, Big Data: The Next Frontier for Innovation, Competition, and Productivity, May 2011.

[3] http://lsst.org/lsst/google, http://www.lsst.org

30 تيرا بايت (30 ألف جيجا بايت) في كل ليلة وذلك يمثل حجما ضخما من المعلومات التي يفترض تحليلها بعد عملية الاسترجاع.

التنوع Variety

تتواجد المعطيات الضخمة على الشابكة في صيغة ملفات متنوعة تتسم بامتدادات مختلفة files extentions يرمز كل امتداد لنوع معين من الملفات (صور، ملفات نصية، قواعد بيانات، تسجيلات فيديو، ... وغيرها من الملفات).

سرعة المعطيات Velocity

ويتعلق الامر بسرعة تنقل المعلومات عبر الشابكة.
قديما، كانت المعلومات تعالج بصفة آلية لكن على مراحل، وذلك لتسمح بمعالجة مقاطع Batch من المعلومات \ من البرنامج بشكل متواصل فكان هناك تحكم في حجم المعلومات الواردة وكذا في الوقت المخصص للمعالجة، غير أن الواقع الرقمي الجديد أوجد طرقا أخرى للمعالجة خاصة في الحالات التي تصل فيها المعلومات تباعا ودون انقطاع.
وقد أصبحت التقانات الحديثة تسمح بمعالجة المعلومات في حين ورودها وبالتالي أصبح من الصعب التنبأ بمدة المعالجة .
ومثال ذلك الموقع الاجتماعي Twitter والذي بلغ معدل التغريدات الواردة في اليوم إلى 140 مليون تغريدة.

2- **البرمجيات المكرسة لتحليل المعطيات الضخمة**

إن أهمية المعطيات الضخمة لا تكمن في حجمها الهائل بل يتعلق الأمر بكيفية تحليلها وتحديد سياقها ومن ثم استغلال المعلومات المستخلصة لاتخاذ قرارات استباقية تتنبأ بوضع معين تتيح تعديله أو تغييره وفي حالات أخرى تسمح على الأقل باتخاذ قرارات مناسبة وتقلل نسبة الخطأ
ويمكن استغلال هذه التقانات في عدة مجالات (علمية- تقنية- اجتماعية وحتى سياسية).

وقد أضحى تحليل المعطيات الضخمة قرارا استراتيجيا تتبناه الهيئات والشركات الكبرى لتحديد التوجهات العامة في سياق معين واستخلاص
نماذج للمعطيات من خلال الكم الهائل المتوفر من المعلومات
ومن بين المقاربات الحديثة المستعملة في تحليل المعطيات الضخمة :

1-2 مقاربة التقسيم والتزامن Divide and concur

وتنتهج المنصة الالكترونية Hadoop[4] هذه المقاربة، إذ يتم تقسيم مجموعات المعطيات الضخمة إلى أجزاء أصغر (HDFS) ومعالجتها (Mapreduce) بطريقة موازية باستخدام الآلاف من المخدمات **Servers** . فمن جهة تسمح هذه الطريقة باحتواء المعطيات الضخمة ومن جهة أخرى ، فإن أسس التقسيم المنتهجة لتصنيف المجموعات تسمح بتدقيق

[4] http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html

المعلومات وبالتالي فهي توفر تفاصيل دقيقة تسمح باستغلالها بطريقة ناجعة، علما بأن هذه المنصة الإلكترونية مصنفة ضمن المصادر المفتوحة.

ويتطلب حجم المعطيات المتضاعف باستمرار زيادة عدد المخدمات ، علما أنه في حال استعمال تقانة الحوسبة السحابية ينقص الضغط على المخدم ويتحسن أداه وبالتالي يتمكن من تقليص مدة المعالجة.

2-2 مقاربة القوة الغاشمة Brute[5] and Force

يشير مصطلح **القوة الغاشمة** إلى نمط البرمجة التي لا تلجأ إلى اختصارمراحل المعالجة بغرض تحسين أدائها وتقليص مدة المعالجة، وفي المقابل تعتمد على قدرة الحاسوب الفائقة على المعالجة والتي تأخذ في الحسبان كل الاحتمالات الممكنة وتجربها واحدة تلو الأخرى حتى التوصل إلى حل للمشكلة.

أما بالنسبة **للمعطيات الضخمة،** فإن انتهاج هذه المقاربة يقتضي استخدام مخدم Server ذو قدرة هائلة على المعالجة وبالتالي يتوفر على ذاكرة ذات حجم هائل يقارب 100 تيرا بايت (علما بأن 1 تيرا بايت يقابل 1000 جيجا بايت)، حيث تتم معالجة المعطيات Data كوحدة واحدة

ويتم ضغط المعطيات في الذاكرة، وقد تصل نسبة الضغط 1:100 عندما يتعلق الأمر بملف نصي صرف. في هذه الحالة بإمكان مخدم يتوفرعلى ذاكرة تبلغ 100 تيرابايت تحميل حزمة من المعطيات الضخمة بحجم الذاكرة وتحليلها ، كما هو الشأن للأداة HANA لتحليل المعطيات الضخمة التي طورتها كل من **شركتي** [6] IBM وSAP.

3- علاقة المعطيات الضخمة بتقانات اللغة

لا يمكن بأي حال من الأحوال عزل المعطيات الضخمة عن اللغة لأنها موجودة أصلا بلغة معينة ، وأيا كان نوع الملفات التي نحن بصدد معالجته فإننا نجد دائما معلومات لغوية مرتبطة به، فمثلا إذا تعلق الأمر بملف يحمل إحصائيات أو معلومات رقمية فإنه حتما يتضمن وسما للمعلومات الرقمية، أما إذا تعلق الأمر بملفات تتضمن صورا فسيتوفر حتما على دليل أو كلمات مفتاحية توضح مضمون الصورة، وكذلك الشأن بالنسبة لتسجيلات الفيديو المرئية أو التسجيلات الصوتية التي يمكن باستخدام تقانات ملائمة تحويل مضمونها إلى ملف نصي وبالتالي يمكننا الربط بين التقانات المسخرة لمعالجة اللغة والمعطيات الضخمة، بل أكثر من ذلك لا يمكن الاستغناء عن هذه التقانات لأنها تمثل مرحلة من مراحل معالجة المعطيات الضخمة.

وعلى سبيل المثال لا الحصر نجد :

- برمجيات التنقيب الآلي واسترجاع المعلومات .
- برمجيات اكتساب المعرفة من الملفات النصية ومواقع التواصل الاجتماعي.
- برمجيات معالجة اللغة للبحث وربط المعلومات

---

[5] http://www.webopedia.com/TERM/B/brute_force.html

[6] http://www.ibm.com/solutions/sap/us/en/landing/hana.html

- برمجيات وسم الملفات غير النصية
- برمجيات تحويل الكلام المنطوق إلى نص وتحويل النص إلى كلام منطوق
- برمجيات تستعمل كواجهة دلالية تسهل عملية التفاعل مع المستخدمين على غرار البرمجيات التي توفر إجابات آلية لطلبات الاستعلام.
- برمجيات توليد ملخصات وتقارير آلية.

وجلها يستعمل في مجال المعالجة الآلية للغات الطبيعية.

4- إيجاد بنية مهيكلة لمعطيات مبعثرة

رغم وجود برمجيات كثيرة لإدارة قواعد البيانات إلا أن أغلب المعلومات التي تمثل المحتوى الرقمي على الشابكة توجد في صيغة معلومات غير مصنفة. فمن مدونات إلى بريد إلكتروني مرورا بتغريدات ووصولا إلى نصوص غير موسومة متناثرة على الشابكة . ورغم ذلك فهي تحمل في طياتها معلومات قيمة يمكن استغلالها شريطة تغيير بنيتها العشوائية .

والبداية تكون من اللغة ، فإذا كانت المعلومات متوفرة بلغة أجنبية نبدأ بترجمتها إلى اللغة العربية.

ثم بعد ذلك يمكن استغلال هذه الترجمة لإنشاء متن لغوي متوازي أو بعبارة أبسط متن لغوي مصفوف على مستوى الجمل بحيث تُصَفُّ كل جملة مع الترجمة المكافئة لها.

من جهة أخرى، يمكن استخراج قائمة المصطلحات من المتن اللغوي، علما أنه هناك خوارزميات تقوم بعملية استخراج المصطلحات من المتن اللغوي بصفة آلية.

كما يمكن استغلال نفس النصوص لإنشاء مسارد .

إن عملية تصنيف المعلومات لا تنتهي عند هذا الحد بل يمكن استغلال كل المراحل السابقة كبداية لمرحلة أخرى من التصنيف، فمثلا بوسعنا تصنيف المصطلحات ضمن قوائم إسمية قياسية مفهرسة ومن ثم إنشاء قواعد للبيانات وبعدها يمكن أيضا أن فهرسة قواعد البيانات في حد ذاتها قاعدة البيانات للتمكن من انشاء قواعد بيانات علائقية.

5- دور الترجمة الآلية في استيعاب المعطيات الضخمة

بديهي أن ترتبط المعطيات الضخمة باللغة التي صيغت بها وبالتالي فالمعلومات القيمة والتحاليل وكل التنبؤات المرتبطة بهذه التحاليل ستكون باللغة التي وجدت عليها هذه المعلومات.

فإذا كانت المعطيات الضخمة موجودة بلغات مختلفة كما هو الحال فعلا، يتعين اختيار اللغة التي نريدها ان تعكس معلوماتنا ونترجم إليها المعطيات قبل البدء في عملية التحليل .

وقد تم تبني هذا الحل فعلا من قبل شركة SDL بحيث أطلقت نظامها الجديد للترجمة الآلية BeGlobal الذي يدعم تقنية الحوسبة السحابية والذي يستعمل منصة إلكترونية جديدة ستتيح ترجمة المواقع الإلكترونية بما فيها مواقع التواصل الاجتماعي ومنصات التجارة الإلكترونية وبذلك ستسمح هذه التقنية بالتنبؤ بالتوجهات التجارية ، فضلا عن إمكانية رصد أحاسيس العملاء

والأفراد وبالتالي رصد أفكار المجتمعات وذلك متاح بثمانين 80 لغة تدعمها المنصة الإلكترونية لـ SDL.

في الوقت التي بلغت فيه نظم الترجمة الآلية في الدول الغربية هذا المستوى لا تزال نظم الترجمة الآلية وبرمجيات المعالجة الآلبة المصممة للغة العربية تعاني غيابا للدعم العربي لمشاريع في هذا المجال وإن وجدت فإنها تمثل مبادرات فردية ولا تعكس مشروعا عربيا.

6- الخاتــمة والتوصيات

لا يزال واقع اللغة العربية بعيدا كل البعد عن الواقع الرقمي الذي أفضى إليه عصر المعلومات فالمعطيات الضخمة التي لم تعد المخدمات العادية قادرة على استيعاب حجمها، واستوجبت استحداث برمجيات جديدة لمواكبتها لم تتمكن من إثراء المحتوى الرقمي العربي الذي لا تفوق نسبته 3%، وعليه يتعين علينا كاختصاصيين التفكير في حلول من شأنها فك العزلة التي تلازم اللغة العربية ومن بين هذه الحلول التي يمكن اقتراحها:

1- تبني سياسة موحدة للبحث العلمي في مجال المعالجة الآلية للغة العربية.
2- دعم الأبحاث العربية في مجالات المعالجة الآلية للغة العربية.
3- إنشاء لجنة علمية لاقتراح مشاريع علمية مكرسة للغة العربية مع تحديد الموارد اللازمة وضمن جدول زمني محدد لاتمامها
4- استغلال المعطيات الضخمة Big Data المتوفرة بكل اللغات وترجمتها وإنشاء برمجيات لتحليل المعطيات الضخمة تدعم اللغة العربية .
5- إنشاء موارد لغوية : مسارد ، مصطلحات تواكب التقانات الحديثة وتوحيد استعمالها .
6- إنشاء برمجيات وموارد لغوية تصنف ضمن المصادر المفتوحة وإتاحتها للباحثين في مجال المعالجة الآلية للغة العربية.

قائمة المراجع

1- Taming the Big Data Tidal Wave : Finding opportunities in HugeData Streams with advanced Analytics **; By Bill Franks**
Edition : 2012 Published by John Wiley & Sons, Inc., Hoboken, New Jersey.Published simultaneously in Canada

2 - Principles of Big Data
Preparing, Sharing, and Analyzing Complex Information ;
By Jules J. Berman, Ph.D., M.D.
Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA
Copyright © 2013 Elsevier Inc **.**

3 - **SDL Turns Multilingual Big Data into Big Text: Integrates Machine Translation with Text Analytics .**
http://www.cmswire.com/cms/customer-experience/sdl-turns-multilingual-big-data-into-big-text-integrates-machine-translation-with-text-analytics-020002.php

4 - Big Data Imperatives: Enterprise 'Big Data' Warehouse, 'BI' Implementations and Analytics
By Soumendra Mohanty, Madhu Jagadeesh, Harsha Srivatsa
Edition : Apress 2013.

5 - MapReduce and MPP: Two Sides of the Big Data Coin,
ZDNet, 2 March 2012

6- **Hadoop Could Save You Money Over a Traditional RDBMS,**
Computerworld UK, 10 January 2012

7- www.almaany.com قاموس إلكتروني عربي - إنجليزي

8- http://searchbusinessanalytics.techtarget.com/feature/In-memory-analytics-tools-pack-potential-big-data-punch