مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(٦) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (6) 2023

# تطوير نظام إرشاد أكاديمي للباحثين باستخدام تقنيات التنقيب في البيانات

**د. راغب طعمه ***

**محمد عيسى علي ****

**(تاريخ الإيداع ٢٠٢٣/٥/١١ . قُبِل للنشر في ٢٠٢٣/٨/١٣ )**

**□ ملخّص□**

يعتبر تحديد واختيار موضوع البحث العلمي بدقة من أهم الأمور إذا لم يكن أهم موضوع في البحث العلمي، ذلك أن تحديد موضوع البحث العلمي بدقة منذ البداية يضمن بداية صحيحة وفعّالة للباحث في رحلة البحث العلمي. أدى تزايد أعداد الجامعات ومراكز الأبحاث بالإضافة إلى الثورة التكنولوجية الهائلة التي شهدها العالم في السنوات الأخيرة إلى إنتاج أعداد ضخمة جداً من المنشورات العلمية في مختلف المجالات الأمر الذي يجعل من الصعب على أي باحث تتبع مسار البحث العلمي بدقة وتحديد التوجهات المستقبلية للبحث. تهدف الدراسة إلى إرشاد ومساعدة الباحثين ضمن تخصص تقانة المعلومات في اختيار مجالات بحث من المرجح ازدهارها مستقبلاً اعتماداً على تحليل المنشورات العلمية في مجال تقانة المعلومات ودراسة تطورها خلال السنوات السابقة بالاعتماد على خوارزميات التنقيب في المعطيات.

يقدم هذا البحث دراسة منهجية لكيفية تحصيل البيانات وتجميعها وتحضيرها لعملية التنقيب ومن ثم إجراء عمليات التنقيب في هذه المعطيات باستخدام مجموعة متنوعة من الخوارزميات التي تحقق الغرض المطلوب، كما يقدم أدوات للبحث واستعراض مجموعات البيانات التي تم تحصيلها من خلال تأمين محرك بحث ضمن المنشورات العلمية التي تقدمها IEEE مع لوحة مراقبة Dashboard لاستعراض بيانات إحصائية أولية عن مجموعة البيانات التي تم تحضيرها.

**الكلمات المفتاحية:** التنقيب في البيانات، نمذجة الموضوعات، استخراج المعلومات، تجميع الوسائل K، تصنيف الأوراق العلمية (IEEE)

* مدرس في قسم هندسة تكنولوجيا المعلومات – هندسة تكنولوجيا المعلومات والاتصالات– جامعة طرطوس–سوريا

** طالب ماجستير في قسم هندسة تكنولوجيا المعلومات–هندسة تكنولوجيا المعلومات والاتصالات–جامعة طرطوس–سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(٦) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (6) 2023

# Developing an Academic Advising System for Researchers by using Data Mining Techniques

**Dr.Ragheb Toemeh ***
**Mohamad Essa Ali ****

**(Received 11/5/2023 . Accepted 13/8/2023)**

□ **ABSTRACT**

Determining and choosing the subject of scientific research accurately is one of the most important things if it is not the most important subject in scientific research, because defining the subject of scientific research accurately from the beginning ensures a correct and effective start for the researcher in the scientific research journey. The increase in the number of universities and research centers, in addition to the massive technological revolution that the world has witnessed in recent years, has led to the production of very large numbers of scientific publications in various fields. This makes it difficult for any researcher to accurately track the course of scientific research and determine future directions for research. The project aims to guide and assist researchers within the field of information technology in selecting areas of research that are likely to flourish in the future, based on the analysis of scientific publications in the field of information technology and studying its development during previous years, based on data mining algorithms.

This research presents a systematic study of how to collect and compile data and prepare it for the mining process, and then conduct mining operations in this data using a variety of algorithms that achieve the desired purpose. It also provides tools for searching and reviewing the data sets that were collected by securing a search engine within the scientific publications provided by IEEE with a Dashboard monitoring panel to view the primary statistical data on the data set that was prepared.

**Key Words:** Data Mining, Topic Modeling, Information Extraction, K-Means Clustering, (IEEE) Paper Classification

* Teacher, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.
** Master student, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

## ١. مقدمة

يعتبر البحث العلمي أهم أداة لمعرفة حقائق الكون والإنسان والحياة في المجالات المختلفة، ويتيح البحث العلمي للباحث الاعتماد على نفسه في اكتساب المعلومات، كما أنه يسمح للباحث الاطلاع على مختلف المناهج واختيار الأفضل منها. حيث أن البحث العلمي هو أسلوب منظم في جمع المعلومات الموثوقة وتدوين الملاحظات والتحليل الموضوعي لتلك المعلومات باتباع أساليب ومناهج علمية محددة بقصد التأكد من صحتها أو تعديلها أو إضافة الجديد لها، ومن ثم التوصل إلى بعض القوانين والنظريات والتنبؤ بحدوث مثل هذه الظواهر والتحكم في أسبابها. كما أنه الوسيلة التي يمكن بواسطتها الوصول إلى حل مشكلة محددة، أو اكتشاف حقائق جديدة عن طريق المعلومات الدقيقة.

إن المجتمعات التي قادت مسيرة الحضارة، هي المجتمعات التي استطاعت أن توظف البحث العلمي على أوسع نطاق، مجتمعات أدركت أن السير الاعتباطي والعفوي لوتيرة الحياة، لا يوصل إلى نتائج محققة، بينما إخضاع الظواهر والمشكلات للدراسة والتحليل، يقود حتماً إلى حلول حتمية أو إلى الحصول على معلومات مفيدة عن المستقبل.

يشهد العصر الحالي تطوراً سريعاً وكثيفاً في البحث العلمي بمجالاته المختلفة، وخاصة في مجال تقانة المعلومات، وظهور طيف واسع جداً من المواضيع الجديدة في هذا المجال، فنحن نعيش في عصر المعلومات والانفجار المعلوماتي وأصبحت تقنيات التعامل مع المعلومات من ضرورات البقاء. وأصبحت المعلوماتية أداة أساسية للبحث العلمي وتنمية المعارف من جهة، وموضوعاً للبحث العلمي من جهة أخرى، وساعد تطور تقاناتها وأدواتها المختلفة من شبكات تناقل البيانات إلى طرق التخزين والتحليل والبحث وصولاً إلى الذكاء الصنعي والنظم الخبيرة وقواعد المعرفة والتنقيب في البيانات واكتشاف المعرفة المخبأة والعديد من التطبيقات المعقدة وبذلك فهي تحتل الصدارة في مجالات البحث العلمي.

أدى هذا التطور إلى خلق مجموعات ضخمة جداً من الوثائق العلمية عبر الزمن والتي تشكل ثروة حقيقية لا تقدر بثمن، وأصبحت عملية الحصول على معلومات مفيدة منها أمر معقد جداً.

كما أدى التطور السريع في البحث العلمي بشكل عام وتطور البحث في مجالات تقانة المعلومات بشكل خاص إلى ظهور مصطلحات جديدة واختفاء مصطلحات أخرى، أو حتى اختفاء مجالات علمية كاملة أو توقف التطور في هذا المجال وذلك بسبب عدم وصول العلماء إلى نظريات جديدة في هذا المجالات. فمثلاً هناك بعض مجالات تقانة المعلومات تعد ساكنة ووصلت إلى مرحلة من التطور لم تعد من المجالات الساخنة التي يبحث فيها العلماء وتوقفت عملية البحث العلمي فيها.

كل ذلك، ساعد على خلق الحاجة إلى التنبؤ بالمواضيع أو المحاور الجديدة التي سوف تأخذ حيزاً مهماً في مجال البحث العلمي في المستقبل والتي تساعد الباحث في اختيار المواضيع التي سيقوم بالتركيز عليها في أبحاثه العلمية المستقبلية.

### ١.١ الأبحاث ذات الصلة

قامت العديد من الأبحاث بتصنيف الأبحاث العلمية المستقبلية باستخدام تقنيات التنقيب في المعطيات منها:

قام A. A. Jalal [1] باقتراح نظام تصنيف أوراق البحث بناءً على تردد المصطلح Term Frequency (TF) وتردد الوثيقة العكسي للتردد Term Frequency–Inverse Document Frequency (TF–IDF) حيث يوفر ترجيح TF–IDF فكرة جيدة عن مدى أهمية الكلمات من خلال ظهور كلمات معينة في محتوى المستندات وتم استخدام تشابه جيب التمام لقياس التشابه بين محتوى المجموعات والمستندات، لتوجيه المستخدمين حسب احتياجاتهم في مجال الأوراق البحثية، وفر النهج المقترح عملية تجميع المستندات التي تعتمد على ثلاثة أجزاء رئيسية من الورقة البحثية وهي العنوان، الملخص والكلمات الرئيسية. أظهرت الخوارزميات المختارة نتائج دقيقة وموثوقة في التصنيف وفقًا لمجموعات محددة مسبقًا حيث أظهرت أنه من الممكن تصنيف أكثر من ٩٦٪ من الأوراق في نطاقات مماثلة باستخدام جيب التمام.

استخدم S. Kim [2] نظام تصنيف أوراق البحث بالاعتماد على تردد المصطلح Term Frequency (TF) وتردد الوثيقة العكسي للتردد Term Frequency–Inverse Document Frequency (TF–IDF) حيث تم تطبيق خوارزمية التجميع K–mean لتصنيف الأوراق ذات الموضوعات المتشابهة، بناءً على قيم Frequency–Inverse Document Term Frequency (TF–IDF) لكل ورقة حيث وفر النهج المقترح عملية تجميع المستندات بالاعتماد على عنصر الملخص ضمن الوثيقة العلمية والكلمات المستخرجة بواسطة مخطط Latent Dirichlet allocation (LDA).أظهرت النتائج التجريبية أن النظام المقترح يمكنه تصنيف الأوراق ذات الموضوعات المتشابهة حسب الكلمات المفتاحية المستخرجة من ملخصات الأوراق حيث يتمتع بأداء أفضل لتجميع الأوراق البحثية.

استخدم N. Arshad [3] نهج جديد لتحديد الاتجاهات العلمية المستقبلية بهدف مساعدة الباحثين من خلال تعدين موضوعات من Call for Papers (CFP) من جهة والمنشورات العلمية المفهرسة في Digital Bibliography Library Project (DBLP) من جهة أخرى حيث تم حساب تردد المصطلح Term Frequency (TF) وتردد الوثيقة العكسي للتردد (TF–IDF) واستخدام نظام تصنيف الحوسبة Association for Computing Machinery Computing Classification System (ACM CCS)، حيث تم تعيين ١١.٣ ألف منشور مفهرس بواسطة DBLP للمؤتمرات ذات الصلة في مجالاتها عن طريق مطابقة الكلمات الرئيسية التي تظهر في عنوان الوثيقة العلمية. أظهرت النتائج ارتفاع "تحليلات البيانات الضخمة" في موضوعات CFP في السنوات الأخيرة؛ في المقابل، تُظهر موضوعات مثل "الويب الدلالي" و "اكتشاف التسلل" انهياراً، كما أكدت نتائج الدراسة أن مؤتمرات المستوى الأعلى لا تحدد بالضرورة اتجاهات البحث، تبين أيضًا أن تحليل الاتجاهات العلمية باستخدام مجموعات بيانات CFP يمكن أن يكون طريقة أفضل من شأنها أن تساعد الباحثين في بداية حياتهم.

قام M. Haq [4] بتحليل مجالات الحوسبة السحابية والبيانات الضخمة للعثور على أحدث الاتجاهات والموضوعات وترتيبها حسب الأهمية حيث يتم تطبيق تقنيات التنقيب عن النص المتنوعة مثل تحليل تردد المصطلح وتحليل التشابه وتحليل الكتلة ونمذجة الموضوع Latent Dirichlet allocation (LDA). باستخدام تحليل تكرار المصطلح وجدنا الكلمات عالية التردد في الأوراق البحثية والكلمات مرتبطة في كلتا فئتي المقالات. يوضح تحليل التشابه من حيث الفئة أن هذه المقالات ليست متشابهة تمامًا ولكنها مترابطة في المعنى في سياق مجالاتها. المقالات في الفئة الأولى أكثر تشابهًا مقارنة بمقالات الفئة الثانية، تُظهر تقنية تحليل الكتلة أن المستندات شديدة الارتباط متوضعة في مجموعة واحدة، فهذا يعني أن هذه المقالات تناقش نفس الموضوع. تقوم تقنية نمذجة الموضوع بتجميع الأوراق في موضوعات مرتبطة منطقيًا. من خلال تحليل الكلمات الرئيسية في الموضوعات وتطبيق خوارزمية التجميع K–mean لتصنيف الأوراق ذات الموضوعات المتشابهة تم تصنيف مجالات البيانات الضخمة العشرة في مقالات الفئة ١ التي يتم فيها استخدام تقنية الحوسبة السحابية. بينما في مقالات الفئة الثانية، اكتشفت هذه الدراسة أربعة عشر عاملاً من عوامل تبني السحابة والعقبات في التبني.

قدم S. A. Salloum, et al [5] نهج يقوم على استخدام تقنيات التنقيب عن النص لاستخراج معلومات مثيرة للاهتمام من المقالات التي تم جمعها في مجال التعلم المتنقل باستخدام تقنيات التنقيب عن النصوص لمساعدة الباحثين بمعرفة المحاور المهمة ضمن هذا المجال حيث تم تطبيق تقنية سحابة الكلمات التي تعد من أكثر الطرق استخدامًا لتقديم البيانات النصية بطريقة رسومية حيث يتم تمثيل الكلمات المكتوبة ضمن محتوى منظم حسب ترددها بالإضافة إلى استخدام خوارزمية K–mean لتصنيف الأوراق ذات الموضوعات المتشابهة. أظهرت النتائج أن مصطلح "التعليم" يظهر على أنه مركزي في هيكل الشجرة حيث يتم توصيل جميع الكلمات ذات الصلة به. يليه "المرضى" و "الطلاب" على التوالي حيث يمكن الإشارة إلى حقيقة أن النص المكتسب من المقالات البحثية المجمعة يركز بشكل أساسي على مجال التعلم. بالإضافة إلى ذلك، تم إجراء مقياس التشابه على المقالات التي تم جمعها من أجل تحديد الموضوعات المتشابهة إلى حد كبير مع بعضها البعض. كشفت النتائج أن عامل التشابه لم يتمكن من اكتشاف تشابه واضح بين بعض الموضوعات، والسبب في أن هذه الموضوعات مترابطة ومتشابهة في المعنى مع بعضها البعض (أي أن جميع المقالات تناقش موضوع التعلم المتنقل في التعليم العالي). تم استخدم خوارزمية k–mean من خلال استخدام قيم k values المختلفة. أشارت النتائج إلى وجود ست مجموعات. تم تجميع جميع المقالات تقريبًا (N = 285) في مجموعة واحدة؛ يشير هذا إلى أن هذه المقالات تناقش الموضوع الرئيسي المدروس (أي التعلم المتنقل في التعليم العالي)

## ٢. أهمية البحث وأهدافه

تنبع فكرة الدراسة من الحاجة الملحة إلى معرفة ما سوف يؤول إليه العلم في مجال تقانة المعلومات نتيجة التطور المتسارع الذي يطرأ على محاوره وأقسامه، حتى يتمكن الباحث من تحديد المواضيع الأكثر أهمية والتركيز عليها في أبحاثه لاكتشاف معارف جديدة لم تكن موجودة من قبل أو تطوير معارف موجودة وتحسينها.

ونحن ومن خلال هذه الدراسة نهدف إلى التنبؤ بمستقبل الأبحاث العلمية خاصة في مجال تقانة المعلومات ومعرفة المجالات والمحاور القابلة للتطور وكيفية سير البحث العلمي فيها، وبالتالي الوصول إلى مخطط بياني يوضح كيف ستكون عملية البحث العلمي في الفترة المراد التنبؤ بها بالإضافة إلى المجالات العلمية التي تطورت والتي توقف البحث فيها ومقدار هذا التطور.

## ٣. منهجية البحث

سوف يتم انجاز هذه الدراسة من خلال أربع مراحل أساسية وهي:

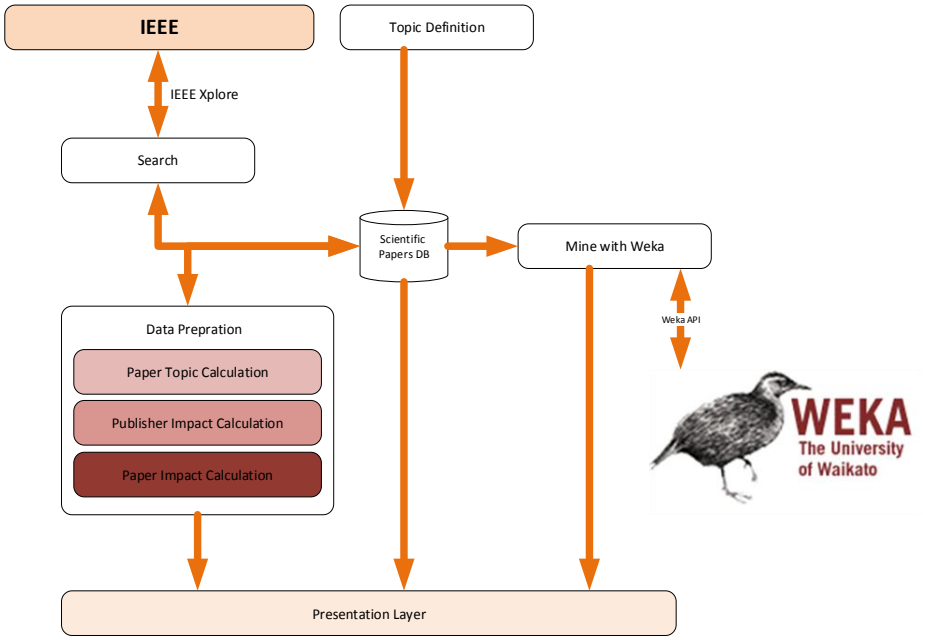
١. بناء برنامج لجمع بيانات الوثائق العلمية المنشورة بعد عام ١٩٩٥ ضمن مواضيع محددة يتم تعريفها من قبل مستخدم البرنامج.

٢. بناء نموذج نظري لتعريف أثر الوثيقة العلمية وربطها بالموضوعات.

٣. جمع وتحضير مجموعة بيانات كبيرة نسبياً وتحضيرها من خلال استكمال البيانات الناقصة وتعريف أثر الوثيقة العلمية وربطها بالموضوعات.

٤. إجراء عمليات تنقيب في مجموعة البيانات التي تم تجميعها باستخدام إحدى أدوات التنقيب في البيانات ودراسة النتائج

يبين الشكل التالي المراحل والمكونات الأساسية لنموذج التنقيب في الوثائق العلمية:




**الشكل (١): المراحل والمكونات الأساسية لنموذج التنقيب في الوثائق العلمية**

يتضمن النموذج عملية تعريف الموضوعات ذات الأهمية للباحث وعملية البحث عن الوثائق وتخزينها في قاعدة بيانات محلية ومن ثم تحضير البيانات عن طريق تعريف موضوعات الوثيقة العلمية وتحديد أثر الوثيقة وفي النهاية الربط مع أداة التنقيب في البيانات Weka لإجراء عملية التنقيب واستحصال النماذج المطلوبة.

## ١.٣ جمع البيانات

### ١.١.٣ تعريف الموضوعات

تم تعريف مجموعة من الموضوعات المتعلقة بمجال علوم الحاسوب حسب الكلمات المفتاحية الأكثر بحثاً في Institute of Electrical and Electronics Engineers (IEEE). وتم ربط كل موضوع بقائمة من الكلمات المفتاحية المعبرة عنه حسب ملخص المصطلحات الخاص بكل موضوع من هذه المواضيع والوارد في مواقع على الانترنت مثل IEEE و W3C و KDNuggets وغيرها.

تجنباً للدخول في درجة تمثيل الكلمات المفتاحية للموضوعات الأمر الذي يحتاج إلى خبير في كل من هذه الموضوعات، تم اعتبار كل كلمة مفتاحية تمثل الموضوع الموافق لها بنسبة ١٠٠%.

تضمنت قائمة الموضوعات المعرفة ما يلي:

(جدول ١): قائمة الموضوعات المعرفة

| Topic | Keywords count |
|---|---|
| Big Data | ١٧٧ |
| Cloud Computing | ٥٧ |
| Data Mining | ٧٧ |
| Internet of Things | ٧٢ |
| Robotics | ٦٢ |
| Semantic Web | ٤٨ |
| Smart Grid | ٤١ |
| Web Services | ٥٧ |

**٢.١.٣ بناء مجموعة بيانات الوثائق حسب الموضوعات المعرفة**

بعد الانتهاء من تعريف الموضوعات، تم استخدام البرنامج لبناء مجموعة بيانات الوثائق العلمية وفق الشروط التالية:

١. إجراء عملية بحث عن كل الكلمات المفتاحية المستخدمة في البرنامج
٢. إجراء عملية البحث بين عامي ١٩٩٥ و٢٠٢٢
٣. إجراء عملية البحث بحيث تعيد ٥٠ نتيجة كحد أقصى لكل عملية
٤. يشترط وجود الكلمة المفتاحية في عنوان الوثيقة
٥. تخزين النتائج المعادة من عمليات البحث بشرط عدم تكرار الوثيقة

بنهاية عمليات البحث تجمعت لدينا مجموعة بيانات تضمنت ١٨٠٩٩ وثيقة والتي ستكون أساساً لعمليات تحضير البيانات والتنقيب اللاحقة.

**٢.٣ تحضير البيانات**

**١.٢.٣ استكمال البيانات الناقصة**

تتضمن عملية استكمال البيانات الناقصة الحصول على بيانات الاستشهاد والتحميل الخاصة بالوثائق. تم إنجاز هذه العملية من خلال طلب HTTPRequest للحصول على صفحة HTML الخاصة بالوثيقة العلمية ومن ثم تحليل الملف للحصول على عنصر Json الذي يخزن هذه البيانات فيه. يقوم البرنامج بقراءة المعلومات المطلوبة وتخزينها في قاعدة البيانات. يتم تنفيذ هذه العملية فقط للوثائق التي لم يتم إحضار بيانات الاستشهاد والتحميل الخاصة بها.

بإنجاز هذه العملية تم بناء مجموعة بيانات متكاملة وأصبح من الممكن استكمال عمليات الحساب المطلوبة قبل البدء بعملية التنقيب.

**٢.٢.٣ ربط الوثائق بالموضوعات**

تتضمن هذه العملية ربط الوثائق الموجودة في مجموعة البيانات بالمواضيع المعرفة من خلال البحث عن الكلمات المفتاحية الخاصة بكل موضوع في عنوان الوثيقة والملخص والكلمات المفتاحية وقائمة المصطلحات الخاصة بكل وثيقة وحساب ارتباط الوثيقة بالموضوع وفقاً للوزن المعطى لكل عنصر من عناصر الوثيقة وفق العلاقة (١) وبالتالي اعتبار الموضوع ذو عامل الربط الأعلى هو الموضوع الخاص بالوثيقة.

$$\text{Topic Weight} = W + X + Y + Z \quad \ldots\ldots\ldots\ldots (1)$$

حيث:

- W: عدد مرات ورود الكلمة المفتاحية في العنوان × معامل تثقيل العنوان.
- X: عدد مرات ورود الكلمة المفتاحية في الملخص × معامل تثقيل الملخص.
- Y: عدد مرات ورود الكلمة المفتاحية ضمن قائمة مصطلحات الوثيقة × معامل تثقيل قائمة مصطلحات الوثيقة.
- Z: عدد مرات ورود الكلمة المفتاحية ضمن قائمة الكلمات المفتاحية الخاصة بالوثيقة × عامل تثقيل قائمة الكلمات المفتاحية للوثيقة.

عندئذ يتم اعتبار الموضوع ذو عامل الربط الأعلى هو الموضوع الخاص بالوثيقة من خلال max (Topic Weight).

تم إنجاز هذه العملية حسب الأوزان المبينة في الجدول(٢):

**جدول (٢): قائمة الأوزان لكل وثيقة**

| العنصر | الوزن |
|---|---|
| العنوان | ١ |
| الملخص | ٠.٦٥ |
| الكلمات المفتاحية | ٠.٧٥ |
| قائمة المصطلحات | ٠.٨ |

تم اختيار هذه الأرقام بناء على أهمية كل عنصر ومدى تمثيله للوثيقة العلمية.

ويتم حساب درجة الترابط باستخدام العلاقة (٢) :

$$\text{Topic Relevance} = 1 - 1/(1 + \max(\text{Topic Weight})) \quad \ldots\ldots\ldots \quad (2)$$

تفيد العلاقة (٢) بحصر درجة ارتباط وثيقة بأحد الموضوعات بقيمة تتراوح بين ٠ و ١ محافظاً على درجة تزايد مقبولة حسب قيمة max (Topic Weight).

بعد إجراء عملية الحساب تبين لدينا توزع الوثائق الموجودة في مجموعة البيانات على الموضوعات وفق الجدول(٣):

**جدول (٣): توزع الوثائق الموجودة في مجموعة البيانات على الموضوعات**

| الموضوع | عدد الوثائق |
|---|---|
| Big Data | ٤٣٣٣ |
| Data Mining | ٤٢٦٤ |
| Internet of Things | ٨٤٦ |
| Robotics | ٢٤٩٢ |
| Smart Grid | ٧١٥ |
| Web Services | ٣٣٨٢ |
| Semantic Web | ١٧١٨ |
| Cloud Computing | ٣٤٩ |

كما تم إنشاء جدول جديد لتخزين معامل ارتباط الوثيقة بكل موضوع من الموضوعات وبالتالي يمكن معرفة مدى قرب الوثيقة التي تنتمي إلى موضوع ما من موضوع آخر، بمعنى آخر معرفة فيما إذا كانت الوثيقة تمثل موضوعين في وقت واحد بدرجة متقاربة.

لم يتم استخدام هذه المعلومة في سياق الدراسة الحالية وإنما تم تركها لتطويرات مستقبلية.

**٣.٢.٣ حساب أثر الناشر**

تم حساب أثر الناشر وفق عدد المنشورات الخاصة بكل ناشر ضمن مجموعة البيانات، ولكن على اعتبار أن مجموعة البيانات تم تجميعها من موقع واحد وهو IEEE فكان من المتوقع أن تكون أغلب المنشورات من منشورات IEEE.

$$\text{Publisher Impact} = 1 - 1000/(\text{sum}(\log(PC_y) * (y - 1994))) \quad \ldots\ldots\ldots \quad (3)$$

حيث تكون $PC_y$ هي عدد المنشورات للناشر في السنة y.

تم توزيع المنشورات على الناشرين وفق الجدول(٤):

**جدول (٤): توزيع المنشورات على الناشرين**

| الناشر | عدد الوثائق | أثر الناشر |
|---|---|---|
| IEEE | ١٧٤٠١ | ٠.٩٩٧١٤٩٤٢٩٧٤٣٤٢ |
| BIAI | ١٠ | ٠ |
| Alcatel-Lucent | ٢ | ٠ |
| IBM | ٢٥ | ٠ |
| AGU | ٣٠ | ٠ |
| Wiley-IEEE Press | ٣١ | ٠ |
| IET | ٤٦٧ | ٠.٨٧٢٦٦٠١٢٩٨٨٦٦٦٨ |
| MIT Press | ٩٢ | ٠.٣٣٣٣٣٣٣٣٣٣٣٣٣٣ |
| VDE | ٣٠ | ٠ |
| SMPTE | ١١ | ٠ |

نلاحظ وجود انحياز واضح جداً لصالح منشورات IEEE الأمر الذي يجعل أثر الناشر تقريباً عديم القيمة في حالتنا هذه، لكن هذه المعلومة تصبح قيمة جداً في حال بناء مجموعة البيانات من مصادر مختلفة.

**٤.٢.٣ حساب أثر الوثيقة**

تم حساب أثر الوثيقة وفق العلاقة (٤) وباستخدام القيم المستخدمة لوزن المعاملات وفق الجدول (٥):

$$\text{Paper Impact} = \log(Y + Cit + Dwn + PI)/(3 + \log(Y + Cit + Dwn + PI)) \quad \text{.....(4)}$$

- Y: سنة النشر × معامل تثقيل سنة النشر.
- Cit: عدد مرات الاستشهاد × معامل تثقيل عدد مرات الاستشهاد.
- Dwn: عدد مرات التحميل × معامل تثقيل عدد مرات التحميل.
- PI: أثر الناشر × معامل تثقيل أثر الناشر.

تم اختيار التابع (٤) المبين أعلاه للحفاظ على درجة تزايد منتظمة للتابع حسب قيمة (Y + Cit + Dwn + PI).

جدول (٥): اوزان المعاملات المستخدمة لحساب أثر الوثيقة

| العنصر | الوزن |
|---|---|
| سنة النشر | ٠.٧٥ |
| عدد مرات الاستشهاد | ١ |
| عدد مرات التحميل | ٠.٩ |
| أثر الناشر | ٠.١ |

تم الحساب وفق هذه الأوزان للمعاملات بناء على أهمية كل معامل. نلاحظ محاولة تجاهل أثر الناشر من خلال إعطائه قيمة صغيرة بينما تم إعطاء عدد مرات الاستشهاد أكبر قيمة كون هذا المعامل يؤثر بشكل مباشر في عملية البحث العلمي.

### ٥.٢.٣ استعراض مجموعة البيانات بعد التحضير

بعد تعريف أثر الوثيقة ودرجة ارتباطها بالمجال أو الموضوع الذي تنتمي إليه يمكن تطبيق مجموعة من التوابع الإحصائية أو تطبيق عمليات التنقيب في البيانات لاكتشاف معدل نمو مجالات البحث العلمي المعرفة.

قمنا بتوفير كلا الأسلوبين لتتبع نمو مجالات البحث العلمي:

١) استخدام التوابع الإحصائية: تم استخدام تابع إحصائي بسيط وفق العلاقة (٥):

$$TopicProgress_y = sum_y(Paper\ Impact * Topic\ Relevance) \quad \ldots\ldots\ldots (5)$$

حيث $sum_y$ هو مجموع جداء أثر الوثائق المنشورة في السنة y في درجة ارتباطها بالموضوع

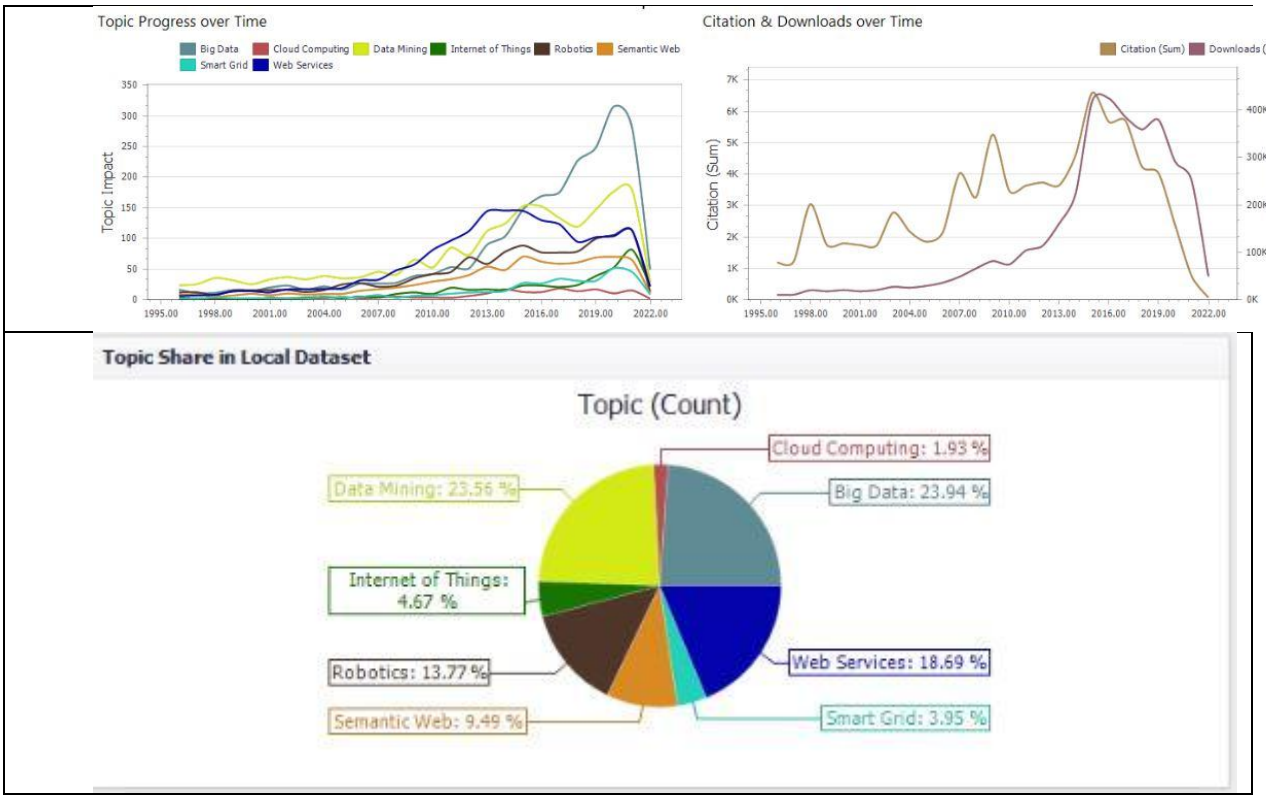
كما قمنا بتعريف تابع النمو النسبي لمجالات البحث العلمي وفق العلاقة (٦):

$$RelativeTopicProgress_y = sum_y(Paper\ Impact * Topic\ Relevance)/count_y(topic) \quad \ldots (6)$$

الذي يأخذ بعين الاعتبار عدد المنشورات المرتبطة بالموضوع خلال سنة ما.

٢) استخدام عمليات التنقيب في البيانات: تم استخدام عمليات التنقيب في البيانات التي توفرها الأداة Weka.

بعد الانتهاء من عملية تحضير البيانات أصبح بالإمكان استعراض مجموعة البيانات الموجودة لدينا وتتبع نمو كل موضوع من الموضوعات المعرفة. يبين الشكل (٢) صورة عامة عن مجموعة البيانات من خلال لوحة المراقبة التي تم تطويرها.




**الشكل (٢): صورة عامة عن مجموعة البيانات من خلال لوحة المراقبة**

يمكن استخلاص القراءات التالية من خلال الشكل(٢):

- تعتبر موضوعات البيانات الضخمة والتنقيب في المعطيات وخدمات الويب الأكثر تمثيلاً ضمن مجموعة البيانات بنسب ٢٣.٩٤% و٢٣.٥٦% و١٨.٦٩% على التوالي

- نلاحظ النمو التدريجي لكافة الموضوعات بتقدم السنين بسبب اعتماد المبدأ التزايدي في حساب نمو الموضوعات. لكن من الملاحظ أن بعض الموضوعات مثل خدمات الويب قد بلغت ذروة نموها بين عامي ٢٠١٢ و٢٠١٥ ومن ثم بدأت بالانحدار التدريجي لكنها حافظت على موقعها كثالث أهم موضوع بعد البيانات الضخمة والتنقيب في البيانات.

- نلاحظ تفوق البيانات الضخمة على بقية الموضوعات وخصوصاً في السنوات الأخيرة مما يدل على النمو المتزايد لهذا الموضوع

- نلاحظ الارتباط بين نمو البيانات الضخمة والتنقيب في البيانات كون هذين الموضوعين مرتبطين عضوياً في حقيقة الأمر.

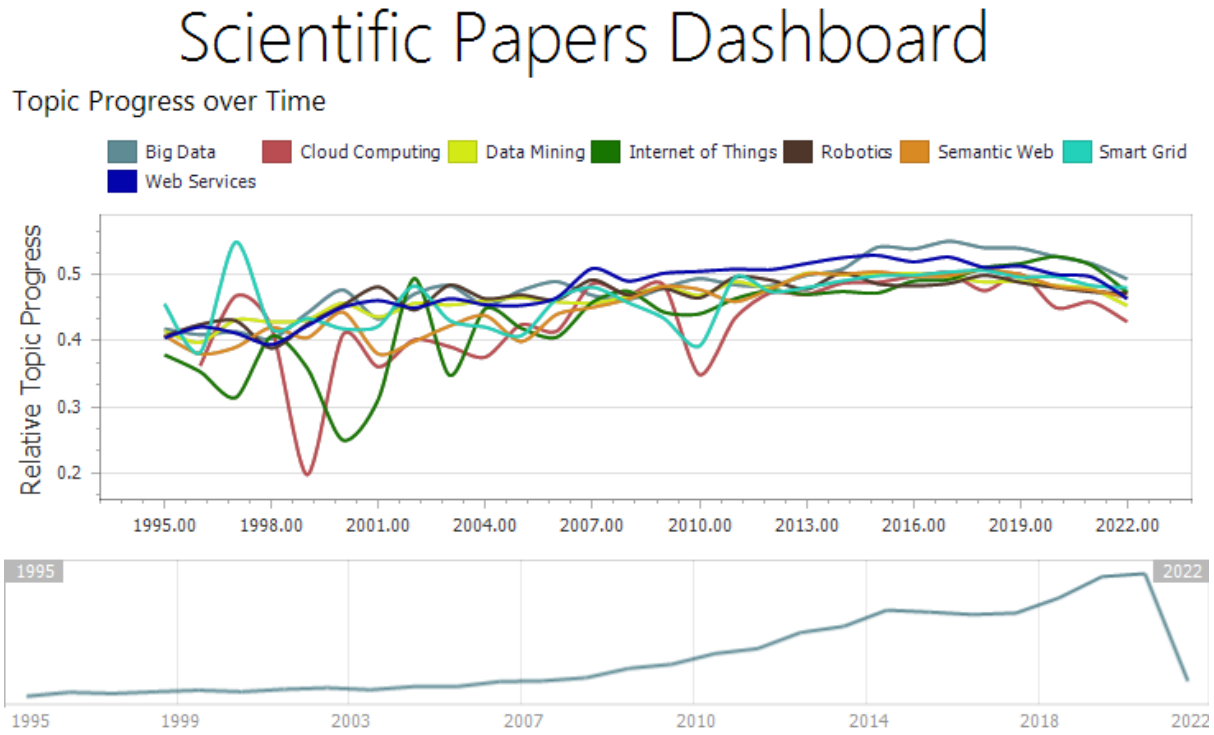
- نلاحظ نمواً متزايداً لموضوع الروبوتيك وحتى تجاوزه لموضوع خدمات الويب في سنة ٢٠٢٠

- بالنسبة لبقية الموضوعات نلاحظ نمواً تدريجياً لكن نسبة أهمية هذه الموضوعات تبقى أقل من غيرها وفقاً لمجموعة البيانات الموجودة لدينا وهذا عائد إلى العدد القليل من الوثائق التابع لكل موضوع من هذه الموضوعات وربما لا يعكس في حقيقة الأمر الأهمية الحقيقية لهذه الموضوعات. يحتاج هذا الأمر لمزيد من الدراسة لمقارنة الموضوعات بالنسبة لعدد الوثائق المتوفرة في كل سنة من السنوات الأمر الذي يؤمنه تابع التقدم النسبي لمجالات البحث العلمي.

- نلاحظ وجود ذروة في عدد مرات الاستشهاد وعدد التحميلات بحدود عام ٢٠١٥ كما توجد عدة ذرى محلية لعدد مرات الاستشهاد خلال أعوام ١٩٩٧ و٢٠٠٣ و٢٠٠٦ و٢٠٠٩. ترتبط هذه النتائج بمجموعة البيانات المتوفرة لدينا والتي تمثل الموضوعات المعرفة وفق نتائج البحث التي يعيدها IEEE.

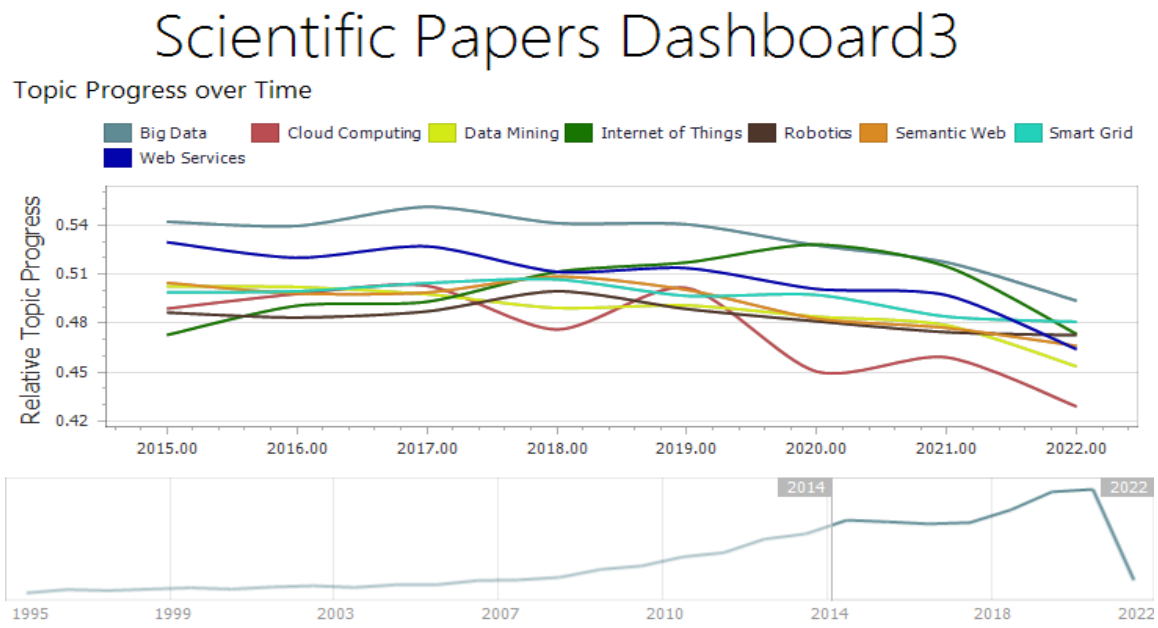
- نلاحظ الانحدار في عام ٢٠٢٢ لكافة المواضيع المحددة لأن عملية جمع البيانات تمت في الفترة الممتدة بين (١٩٩٥-٢٠٢٢).

من خلال استعراض لوحة مراقبة النمو النسبي لمجالات البحث العلمي الشكل(٣) نلاحظ وجود تذبذب في نمو كل من مجالات الحوسبة الشبكة الذكية والحوسبة السحابية وانترنت الأشياء بينما تشهد بقية المجالات نمواً مضطرداً.




**الشكل (٣): لوحة مراقبة النمو النسبي لمجالات البحث العلمي**

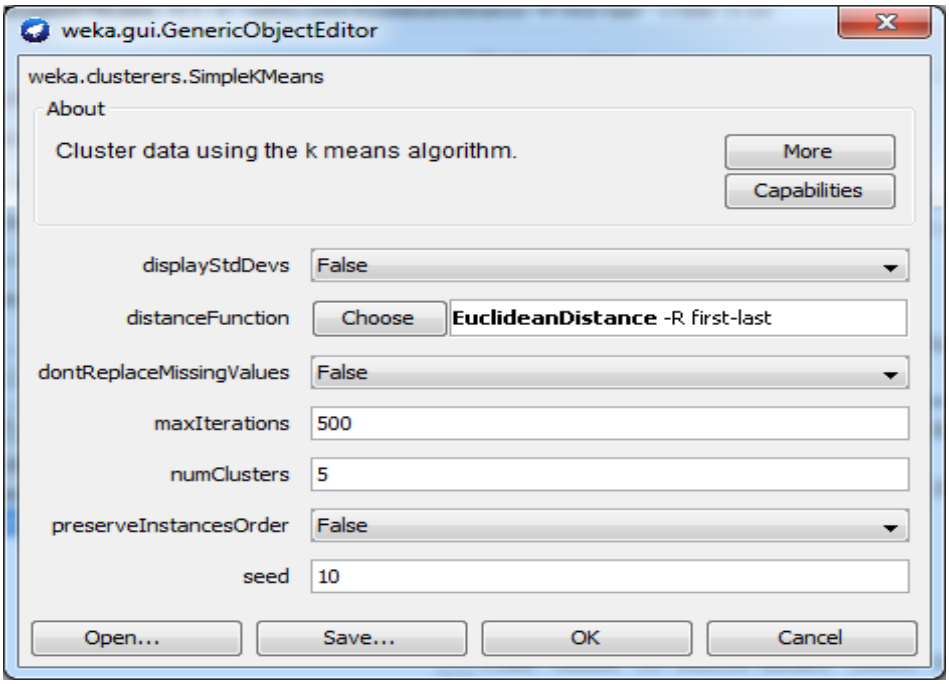
بالنظر إلى الفترة بين أعوام ٢٠١٥ و٢٠٢٢ الشكل(٤) نلاحظ تفوق مجال البيانات الضخمة على بقية المجالات ونمو كبير لأنترنت الأشياء مقابل انحدار لمجال خدمات الويب كما نلاحظ النمو المتزايد للبيانات الضخمة على حساب التنقيب في البيانات على اعتبار أن هذين المجالين مرتبطين ارتباطاً وثيقاً وهذا أمر طبيعي بسبب التركيز الكبير في السنوات الأخيرة على البيانات الضخمة.




**الشكل (٤): لوحة مراقبة النمو النسبي للفترة بين أعوام ٢٠١٥ و٢٠٢٢**

### ٣.٣ التنقيب في قاعدة بيانات الوثائق العلمية

قمنا باستخدام أداة التنقيب في البيانات WEKA لتنفيذ خوارزميات التنقيب في البيانات على مجموعة بيانات الوثائق العلمية .من الواضح أن عملية تتبع نمو مجالات البحث العلمي تفرض إجراء التنقيب في مجموعة جزئية من الأعمدة المكونة لقاعدة البيانات وهي (الموضوع –عدد مرات الاستشهاد –عدد مرات التحميل –أثر الورقة العلمية) تم تصدير بيانات هذه الأعمدة إلى ملف بصيغة Comma Separated Values (csv) وهي من الصيغ التي تتعامل معها WEKA بشكل مباشر. إن المسألة المطروحة أمامنا وهي تتبع نمو مجالات البحث العلمي وأثرها تستدعي إجراء عملية التنقيب باستخدام خوارزميات العنقدة مثل K–Means تم تنفيذ هذه الخوارزمية على مجموع البيانات وفق المعاملات المبينة في الشكل(5) :




**الشكل (٥): خوارزميات العنقدة K–Means على مجموع البيانات وفق المعاملات المبينة**

تم توزيع البيانات على خمسة عناقيد الجدول (٦)، فيما يلي النقطة المركزية لكل عنقود:

جدول (٦): توزيع البيانات على العناقيد

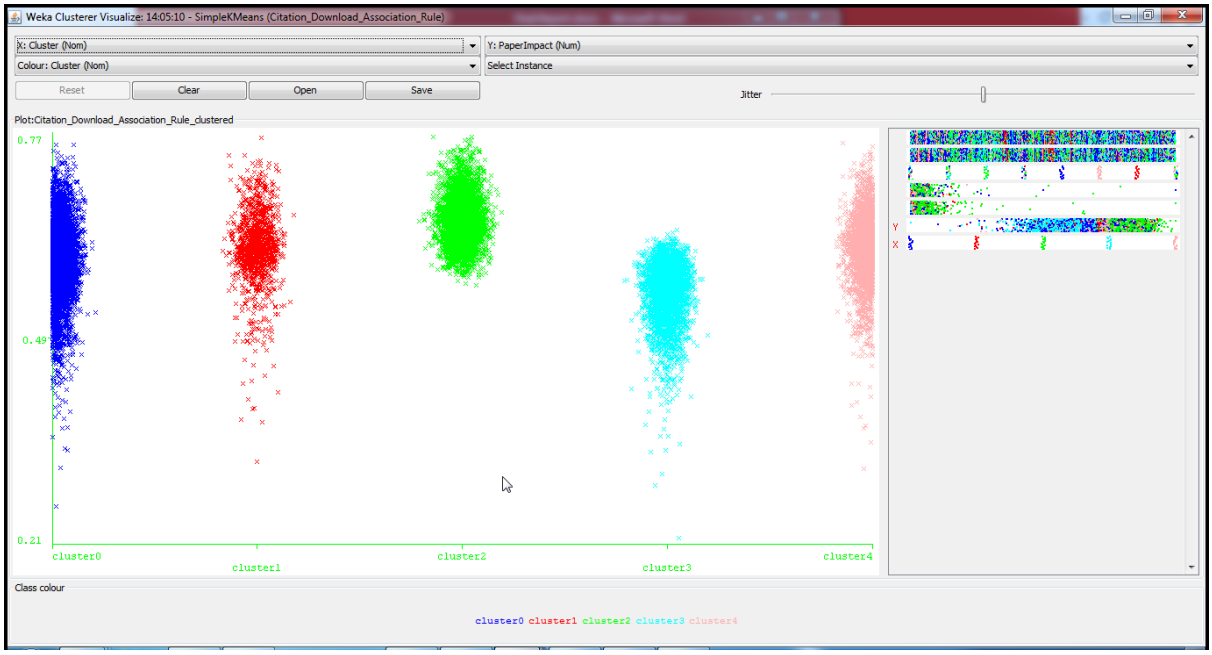
| Cluster | Full Data | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Instnces | 18099 | 5780 | 1298 | 4233 | 3322 | 3467 |
| Percentage | 100% | 32% | 7% | 23% | 18% | 19% |
| Topic | Big Data | Data Mining | Internet of Things | Big Data | Big Data | Robotics |
| Citations | 4.5922 | 3.892 | 3 | 9.1755 | 0.8898 | 4.3066 |
| Downloads | 202.2592 | 132.0965 | 208.9136 | 441.4765 | 35.776 | 184.1915 |
| PaperImpact | 0.6055 | 0.5952 | 0.614 | 0.6508 | 0.5524 | 0.6148 |

يتضمن الجدول:

السطر الأول يمثل العناقيد التي تم توزيع الوثائق لها، السطر الثاني يمثل الوثائق التي تم توزيعها ضمن كل عنقود، السطر الثالث يمثل نسبة تمثيل قاعدة البيانات الكلية من الوثائق لكل عنقود، السطر الرابع يمثل موضوع كل عنقود، السطر الخامس يمثل عدد مرات الاستشهاد للوثائق ضمن العنقود الواحد، السطر السادس يمثل عدد مرات التحميل للوثائق ضمن العنقود الواحد، السطر السابع يمثل أثر الوثائق العلمية لموضوع ضمن العنقود الواحد.

سنقوم بإجراء عملية تحليل للنتائج باستخدام أدوات إظهار البيانات التي تقدمها WEKA.

أولا: توزع العناقيد حسب أثر الورقة العلمية




**الشكل (٦): يظهر الشكل السابق توزع العناقيد وارتباطها بأثر الورقة العلمية**

يمكن استخلاص القراءات التالية من الشكل(٦):

-إن الوثائق التي تنتمي إلى العنقود رقم ٢ ذات تأثير كبير ومرتفع في سياق البحث العلمي وهذا العنقود تنتمي معظم نقاطه إلى موضوع Big Data.

- إن الوثائق التي تنتمي إلى العنقود رقم ١ ذات تأثير متفاوت وبتوزع منتظم في سياق البحث العلمي وتنتمي معظم نقاطه إلى موضوع Data mining وهذا يدل على أن هذا الموضوع يحتفظ بدرجة من الأهمية لعلاقة الوثيقة بمواضيع هامة أخرى

- إن الوثائق التي تنتمي إلى العنقود رقم ٤ ذات أثر كبير نسبيًا لكنها تأتي خلف الوثائق التي تنتمي إلى العنقود رقم ٢ حيث تنتمي معظم نقاط العنقود رقم ٤ إلى موضوع Robotics.

- إن الوثائق التي تنتمي إلى العنقود رقم ٣ ذات تأثير منخفض نسبياً وتنتمي معظم نقاط هذا العنقود إلى مواضيع Semantic Web،Web Service، Big Data. هذا يدل على أنه بالرغم من أن موضوع Big Data ذو أثر مرتفع كما ذكرنا سابقاً إلا أنه توجد مجموعة لا بأس بها من الأبحاث في هذا المجال لا تؤثر في البحث العلمي بشكل كبير مما يدل على وجود توجه لدى عدد كبير من الباحثين للبحث والنشر في هذا الموضوع نظراً لأهميته وانتشاره في الوقت الحالي.

– إن الوثائق التي تنتمي إلى العنقود رقم ١ ذات تأثير متوسط وهذا العنقود ذو كثافة منخفضة حيث تنتمي معظم نقاط هذا العنقود إلى موضوع Internet of Things.

يبين الجدول(٧) مقارنة بين الدراسات المرجعية والنظام المقترح مع التوضيح أن كل الدراسات أعطت نتائج صحيحة ودقيقة حسب الآلية المستخدمة والغرض المطلوب من الدراسة لكن مع مرور الزمن وتقدم التكنولوجيا ظهرت بعض السلبيات فكان لابد من تطوير نموذج ديناميكي يتفادى الوقوع بها:

**جدول (٧): مقارنة الدراسات المرجعية مع النظام المقترح**

| الدراسة | السلبيات | الايجابيات |
|---|---|---|
| **الدراسة الأولى** | – عدم وجود ديناميكية في عملية التنقيب حيث تم التنقيب على مستوى الملخص فقط ضمن الوثيقة العلمية.<br>– لا يتيح النظام للمستخدم أي إمكانية بإدخال أي بيانات جديدة. | – نظام بسيط للتنقيب بالبيانات يعمل ضمن أساليب محددة مما يقلل من زمن الاستجابة.<br>– يوفر دراسة نظرية للباحث للتقدم ضمن مجال البحث العلمي. |
| **الدراسة الثانية** | –التركيز على عنصر الملخص فقط ضمن الوثيقة العلمية.<br>–عدم وجود ديناميكية في عملية تحضير البيانات.<br>–استخدام قواعد بيانات جاهزة (يوجد بارامترات مهمة تتغير بمرور الزمن كعدد مرات التحميل والاستشهاد وأثر الوثيقة العلمية تعتبر أساسية لتتبع المجالات بمرور الزمن). | – فعالية النظام المستخدم عند تصنيف الموضوعات المتشابهة حسب الكلمات المفتاحية المستخرجة من ملخصات الأوراق.<br>–يوفر دراسة نظرية للباحث للتقدم ضمن مجال البحث العلمي. |
| **الدراسة الثالثة** | –التركيز على عنوان الوثيقة العلمية في عملية التنقيب.<br>–افتقار الدراسة الى آلية ديناميكية لجمع البيانات حيث تم استخدام قواعد بيانات جاهزة على مواقع الانترنت.<br>– لا يتيح النظام للمستخدم أي إمكانية للتعديل. | –فعالية النظام المستخدم عند تصنيف الموضوعات المتشابهة حسب الكلمات المفتاحية المستخرجة من عنوان الوثيقة العلمية.<br>–يوفر دراسة نظرية للباحث للتقدم ضمن مجال البحث العلمي. |
| **الدراسة الرابعة** | –عملية التنقيب بالاعتماد على عنصر الملخص ضمن الوثيقة العلمية.<br>–افتقار الدراسة الى آلية ديناميكية لتحضير البيانات. | –فعالية النظام المستخدم عند تصنيف الموضوعات المتشابهة حسب الكلمات المفتاحية المستخرجة من ملخصات الأوراق.<br>–يوفر دراسة نظرية للباحث للتقدم ضمن مجال البحث العلمي. |
| **الدراسة الخامسة** | –عملية تحضير البيانات تقتصر على حساب تردد الكلمات المفتاحية فقط.<br>– لا يتيح النظام للمستخدم أي إمكانية للتعديل. | –استخدام العنوان والملخص ضمن عملية التنقيب مما يوفر آلية متقدمة لاستخراج المعلومات.<br>–يوفر دراسة نظرية للباحث للتقدم ضمن مجال البحث العلمي. |
| **النظام المقترح**<br><br>**النظام المقترح** | –عدم استخدام تقنيات التحليل الدلالي عند كتابة الكلمات المفتاحية الخاصة بالموضوعات.<br>–جمع البيانات بالاعتماد على مصدر واحد فقط (IEEE). | – يتيح للمستخدم استخدام البرنامج بشكل ديناميكي وكامل بكافة مراحل التنقيب من عملية جمع البيانات وتحضيرها الى عملية التنقيب النهائية.<br>– التنقيب على مستوى بنية الوثيقة العلمية (استخدام العناصر الأساسية ضمن الوثيقة العلمية العنوان + الملخص + الكلمات المفتاحية + المصطلحات) كما تتضمن عملية تحضير البيانات العناصر الفرعية ضمن الوثيقة العلمية من (عدد مرات الاستشهاد والتحميل وسنة النشر ...إلخ).<br>–إجراء عملية التنقيب باستخدام توابع احصائية وأداة Weka.<br>–عملية جمع البيانات تتم من خلال توظيف محرك بحث IEEE ضمن النظام بحيث يقوم مستخدم النظام ببناء قاعدة بيانات ديناميكية (قابلة للتعديل على البيانات مع مرور الزمن) خاصة لعملية التنقيب. |

## ٤. خلاصة وتوجهات مستقبلية

### ١.٤ خلاصة

لقد تمكنا في هذه الدراسة من تقديم مقدمة نظرية شاملة عن مفاهيم التنقيب في البيانات وأدواته وتطبيق هذه الدراسة على مسألة هامة وهي تتبع نمو مجالات البحث العلمي عبر الزمن.

في هذا الدراسة تم تغطية كافة مراحل التنقيب في البيانات من عملية جمع البيانات وتنظيفها وتحليلها واستخدامها في مسائل التنقيب التي تم تعريفها لاحقاً. حيث قمنا بجمع عينة بيانات ضخمة نسبياً تشمل بيانات أكثر من ١٨٠٠٠ وثيقة علمية تم الحصول عليها من موقع IEEE ومن ثم قمنا بدراسة هذه البيانات من خلال تطبيق مجموعة من التوابع لتعريف الموضوعات وحساب أثر الناشر وأثر الورقة العلمية من خلال نموذج نظري شامل لكافة مراحل العملية.

يتضمن النموذج المقدم في الدراسة: تعريف الموضوعات والكلمات المفتاحية وربط الوثائق العلمية بالموضوعات من خلال تابع إحصائي تم تعريفه لهذا الغرض وبعد ذلك يتم حساب أثر الناشر وأثر الورقة العلمية باستخدام توابع إحصائية أيضاً.

بعد تحضير البيانات يتم دراسة نمو مجالات البحث العلمي بطريقتين:

- إحصائية: تتم باستخدام تابع نمو مجالات البحث العلمي المطلق وتابع نمو مجالات البحث العلمي النسبي
- طرائق التنقيب في البيانات: تم تعريف عدة مسائل للتنقيب وحلها باستخدام أداة التنقيب في البيانات Weka.

أظهرت النتائج المستخلصة من هذه الدراسة فعالية وديناميكية النموذج النظري المقدم بحيث يتيح لمحلل البيانات إمكانية تحكم عالية المستوى بعملية جمع البيانات وتحضيرها والتنقيب فيها.

تم تطوير البرنامج الخاص بهذه الدراسة باستخدام بيئة العمل visual studio 2015 مستخدمين لغة البرمجة C# وقواعد بيانات SQL Server 2014. كما تم استخدام أداة ال Weka للتنقيب في البيانات وفق الغرض المطلوب.

### ٢.٤ توجهات مستقبلية

إن النموذج النظري المقدم بشكله الحالي يعتبر جيداً نسبياً كنموذج إحصائي لكنه يفتقد إلى تطبيق تقنيات التنقيب في النصوص المتقدمة والتحليل الدلالي لموضوعات البحث العلمي، لذلك يجب في المستقبل العمل على تطوير هذا النموذج من خلال دمج واستخدام تقنيات التحليل الدلالي والبيانات المترابطة والتنقيب في النصوص بهدف تعريف الموضوعات بشكل أدق وباستخدام توابع حساسة للغة أكثر من التوابع الإحصائية المستخدمة في النموذج الحالي.

كما يجب العمل على توفير آلية لجمع البيانات من عدة مصادر بدلاً من الاعتماد على IEEE فقط بهدف بناء مجموعات بيانات شاملة. كما يجب العمل على تعريف الموضوعات والكلمات المفتاحية بناء على تصنيفات قياسية معتمدة الأمر الذي واجهنا صعوبة في تأمينه في المشروع الحالي حيث تم الاعتماد على تصنيفات من مصادر متعددة قياسية وغير قياسية.

على صعيد تأمين وظائف أكثر من خلال البرنامج يمكن تطوير البرنامج ليصبح أكثر ديناميكية من خلال تمكين محلل البيانات من تعريف التوابع المستخدمة في البرنامج وبالتالي سيكون البرنامج يتمتع بالديناميكية الكاملة والشاملة.

**قائمة المصطلحات العلمية**

| الاختصار | المعنى باللغة الانكليزية | المعنى باللغة العربية |
|---|---|---|
| LDA | Latent Dirichlet allocation | تخصيص ديريتشليت الكامنة |
| TF-IDF | Term Frequency - Inverse Document Frequency | تردد المصطلح –تردد المستند العكسي |
| DBLP | Digital Bibliography & Library Project | الببليوغرافيا الرقمية ومشروع المكتبة |
| CFP | call for papers | دعوة لإعداد الأوراق |
| ACM | Association for Computing | جمعية آلات الحوسبة |
| XML | Extensible Markup Language | لغة التوصيف القابلة للتوسيع |
| API | Application Programming Interface | واجهة برمجة التطبيقات |
| csv | Comma Separated Values | قيم مفصولة بفواصل |
| Json | JavaScript Object Notation | ترميز الكائنات باستعمال جافا سكريبت |
| html | HyperText Markup Language | لغة توصيف النص الفائق |
| IEEE | Institute of Electrical and Electronics Engineers | معهد مهندسي الكهرباء والإلكترونيات |
| ACM | Association for Computing Machinery | جمعية آلات الحوسبة |
| CCS | Computing Classification System | نظام تصنيف الحوسبة |

## ٥. المراجع

[1] A. A. Jalal, and B. H. Ali, "Text documents clustering using data mining techniques," International Journal of Electrical and Computer Engineering, vol. 11, no. 1, pp. 664–670, 2021. Doi: https://doi.org/10.11591/ijece.v11i1.pp664-670

[2] S. Kim and J. Gil, "Research Paper Classification Systems based on TF-IDF and LDA Schemes," Human-centric Computing and Information Sciences, vol. 9, no. 30, pp. 1-21, 2019.

[3] Arshad, N., Bakar, A., Soroya, S.H., Safder, I., Haider, S., Hassan, S.-U., Aljohani N.R., Alelyani, S. and Nawaz, R.,"Extracting scientific trends by mining topics from, 2019 Call for Papers", Library Hi Tech, Vol. ahead-of-print No. ahead-of-print

[4] M. Haq, Q. Li and S. Hassan, "Text Mining Techniques to Capture Facts for Cloud Computing Adoption and Big Data Processing", IEEE Access, vol. 7, pp. 162254-162267, 2019. Available: 10.1109/access.2019.2950045.

[5] S. A. Salloum, et al., "Using Text Mining Techniques for Extracting Information from Research Articles," Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence, vol. 740, pp. 373-397, 2018.