مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (6) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (6) 2017

Development of GDFS Web Cache Algorithm Using Normalized Google Distance

Dr. Ali Suleiman[*]
IhabAldibaja[**]

(Received 5 / 11 / 2017. Accepted 12 / 12 / 2017)

□ ABSTRACT □

This paper presents a proposed development of the GDFS algorithm for the Web Cache technology by introducing the Normalized Google Distance (NGD) distance to the algorithm's replacement function to determine the semantic similarity between the objects in the cache and thus support the original algorithm replacement decision when specifying the objects which must be evicted from the cache The study was applied on the information systems operating in Lattakia Port, The results showed that the introduction of semantic similarity using NGD raised the Hit Rate compared to the original algorithm as it improved system performance by reducing page loading time compared with the original successor replacement of the algorithm.

Keywords: Web Cache – Information Retrieval Systems- Semantic Web –Greedy Dual.

[*] **Associate Professor,Department of Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia, Syria.**
[**] **Postgraduate Student, Department of Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia, Syria.**

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (6) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (6) 2017

تطوير خوارزمية كاش الويب GDFS باستخدام مسافة غوغل المقيّسة

د. علي سليمان*
إيهاب الديباجة**

(تاريخ الإيداع 5 / 11 / 2017. قُبِل للنشر في 12/ 12 / 2017)

□ ملخّص □

تقدّم هذه الورقة البحثية تطويراً مقترحاً لخوارزمية GDFS (Greedy Dual Frequency Size) الخاصة بتقنية كاش الويب من خلال إدخال مسافة غوغل المقيّسة NGD(Normalized Google Distance) إلى تابع الاستبدال Replacement Function الخاص بالخوارزمية لتحديد التشابه الدلالي Semantic Similarity بين الأغراض الموجودة في الكاش، وبالتالي، دعم قرار تابع الاستبدال الأصلي الخاص بالخوارزمية لدى تحديد الأغراض الواجب إخراجها evicted من الكاش. تمّ تطبيق الدراسة على أنظمة المعلومات العاملة في الشركة العامة لمرفأ اللاذقية، بينت النتائج أنّ إدخال التشابه الدلالي باستخدام مسافة غوغل المقيسة رفع من معدل الإصابة Hit Rate مقارنةً مع الخوارزمية الأصلية، كما تحسن أداء النظام من خلال تخفيض زمن تحميل الصفحة مقارنةً مع تابع الاستبدال الأصلي للخوارزمية.

الكلمات المفتاحية: كاش الويب – نظم استعادة المعلومات – الويب الدلالي –Greedy Dual.

* أستاذ مساعد – قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية– جامعة تشرين – اللاذقية – سورية
**طالب دراسات عليا (دكتوراه) – قسم هندسة الحاسبات والتحكم الآلي – كلية الهندسة الميكانيكية والكهربائية– جامعة تشرين – اللاذقية – سورية

مقدمة

تواجه أنظمة المعلومات نمواً متسارعاً من ناحية حجم المعلومات المخزنة في هذه النظم الأمر الذي يسبب ازدحاماً (Congestion) كبيراً على الشبكة و تحميلاً زائداً (Overloading) على المخدمات. من الملاحظ أن عدداً كبيراً من نظم المعلومات يقوم بتقديم المحتوى والتفاعل مع المستخدمين من خلال صفحات ويب WIS(Web Based Information Systems لما توفره من واجهة رسومية مرنة وكونها مستقلة عن منصة التشغيل Platform Independent ولا تتطلب برمجيات إضافية على حاسب الزبون.[1]

بسبب هذا النمو المتسارع في المحتويات وعدد مستخدمي النظم ولتقليل حجم المعطيات المنقول على شبكات الحاسب، تم تطوير تقنية التخزين المؤقت لكاش الويب Web Caching وهي تقنية من خلالها يتم تخزين أغراض الويب Web Objects (كصفحات الويب و المستندات....) بشكل مؤقت من أجل استخدامها لاحقاً.[2]

مشكلة البحث

قامت الشركة العامة لمرفأ اللاذقية بتطوير مجموعة من نظم المعلومات الخاصة بالعمليات المرفئية والمرتبطة مع بعضها، يتميز العمل في المرفأ بالحاجة إلى أداءٍ عالٍ لنظم المعلومات لضمان سرعة الوصول إلى المعلومات وعدم حصول تأخير في العمل على أرض الواقع خصوصاً في أوقات الضغط، إن آلية العمل في المرفأ تتطلب وصول مكثف إلى معلومات معينة في لحظات محددة (مثلاً، تزايد الوصول إلى معلومات المانيفست الالكتروني لدى البدء بتفريغ السفينة) ويلاحظ حصول بعض التأخير في استعادة المعلومات في لحظات الذروة[3].

نطرح هنا تطبيق خوارزميات كاش الويب كحل لهذه المشكلة. تم إجراء البحث على خوارزمية GDFS الأساسية، كما تم تطوير الخوارزمية وإدخال مفهوم التشابه الدلالي من خلال مسافة غوغل المقيسة ومن ثم تقييم الخوارزمية الجديدة.

أهمية البحث وأهدافه:

أهمية البحث

تأتي أهمية البحث من تقييم أداء التطوير المقترح من خلال قياس معدل الإصابة Hit Rate وزمن تحميل الصفحة Page Load Duration قبل وبعد تطوير الخوارزمية، حيث نفترض أنّ هذه الخوارزمية من خلال المفاتيح الخاصة بتابع الاستبدال (حجم الغرض – تردد طلب الغرض – كلفة المعالج – كلفة الدخل والخرج – التشابه الدلالي) سوف تسمح بالتعرف على نموذج طلب المستخدم للأغراض في نظام المعلومات المدروس بالشكل الأفضل، وبالتالي، السماح للخوارزمية باتخاذ القرار الأفضل لإدخال وإخراج الأغراض من ذاكرة الكاش.

أهداف البحث

هدف البحث هو تطوير خوارزمية GDFS من خلال تمثيل صفحات الويب في النظام المدروس دلالياً Semantic Web على شكل مجموعة أغراض ترتبط مع بعضها، ومن ثم إدخال التشابه الدلالي بين محتوى الأغراض الموجودة في الصفحات باستخدام مسافة غوغل المقيسة إلى تابع الاستبدال للخوارزمية كطريقة جديدة لتحسين تابع الاستبدال، وبالتالي، تحسين أداء نظام المعلومات المدروس حيث تجمع الخوارزمية المقترحة كلاً من حجم الغرض Size والتردد Frequency وكلفة الدخل والخرج I/O Cost وكلفة المعالج CPU Cost للحصول على الغرض، بالإضافة إلى التشابه الدلالي مع الأغراض الموجودة في الكاش.

طرائق البحث ومواده

تمّ بناء مخدم خاص بالكاش يتضمن الوحدات البرمجية التي تسمح بتطبيق تابع الاستبدال لخوارزمية الكاش المدروسة، كما يتضمن المخدم وحدات تخزين لمعلومات وصفيَة metadata سوف تستخدم في دراسة أداء الخوارزمية، تتضمن هذه المعلومات كامل الحركات على ذاكرة الكاش، بالإضافة لتسجيل أزمنة تحميل الأغراض وعدد مرات الوصول إلى الغرض وغيرها من المعلومات. تمّ تقييم نموذج وصول المستخدم ودراسة المتوسطات الحسابية وتمثيل القيم بخطوط بيانية سمحت بمقارنة معدل الإصابة وأزمنة تحميل الصفحات وبالتالي تقدير التحسين المدخل بتطبيق الخوارزمية المدروسة.

عينات البحث

تمّ دراسة حوالي 700000 سجل تتضمن معلومات الوصول إلى الصفحات المسجلة في نظام المانيفست الالكتروني في الشركة العامة لمرفأ اللاذقية قبل وبعد تطوير خوارزمية الكاش، تمّ الحصول على السجلات من كل من سجلات الوصول الخاصة بمخدم نظام المانيفست الالكتروني IIS(Internet Information Service) Log، ومن البيانات الوصفيّة التي قام مخدم الكاش بتخزينها خلال فترة البحث. امتدت فترة البحث خلال عام 2016 والنصف الأول من عام 2017.

الإطار النظري للبحث

1-الويب الدلالي Semantic Web

الويب الدلالي هو رؤية وفكرة مخترع شبكة الويب WWW "تيم برنارز لي Tim Berners-Lee"، ويهدف من ورائها إلى جعل الويب الدلالي وسيطاً عالمياً لتبادل المعلومات والبيانات والمعارف البشرية، طرح "لي" مفهوم الويب الدلالي في عام 2001 كالتالي:

"إنّ الشبكة العنكبوتية الآن بصورتها الحالية مفهومة لنا نحن بني البشر، ولكنها بالنسبة للحاسب عبارة عن صفحات ممثلة بأصفار وواحدات لا تعني له شيء. إن ما نريده من الويب ذات البنية الدلالية أن يجعل الحاسب أيضا يفهم ماذا تعني محتويات الصفحة التي يقوم بعرضها سعيا نحو التكامل للمحتوى في البيئة العنكبوتية"[4]

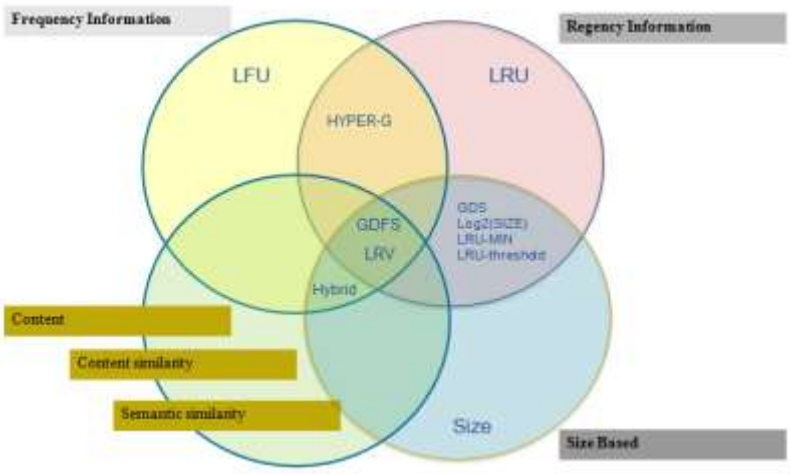
إذاً، يعتبر الويب الدلالي أداةً لتمثيل المعرفة تحصر المصطلحات التي تعبر عن الموضوعات المعرفية والعلمية، وتنظمها موضحة العلاقات المختلفة التي تربط بينها، حيث توضح المصطلحات المترادفة، وذات الصلة، والأعراض، والأضيق، كما تحلل الصيغ المختلفة من المصطلحات، بما في ذلك: الأسماء والأفعال والصفات والظروف وغيرها وذلك لتقديم المعلومات بطريقة يمكن أن يفهمها الحاسب بهدف جعل الحواسيب أكثر ذكاءً.

سوف يفيد هذا بالطبع في عمليات استرجاع المعلومات Information Retrieval من الشبكة العالمية بصفة عامة، وسيعود بالفائدة الكبيرة على قطاع المعلومات الذي سوف يتأثر قطعاً بهذا التقدم في تقنيات استرجاع المعلومات وتقديم خدمات راقية للمستفيدين منها اعتماداً على هذه التقنية.

2-7-كاش الويب Web Cache :

تصنّف خوارزميات الكاش بناءً على ثلاث معايير تحدد سياسة الاستبدال الخاصة بالصفحات وهي الحداثة Recency، وتواتر استخدام الصفحة Frequency ، وقياس الصفحة Page Size. [5] سوف نقوم بإضافة محتوى الصفحة كأحد المعايير التي تحدد سياسة الاستبدال وذلك من خلال تحديد محتوى الصفحة بناءً على تقنيات الويب

الدلالي من أجل دراسة الصفحات الموجودة في الكاش وتقييم علاقة هذه الصفحات مع بعضها من خلال تحديد معايير تشابه بينها بناء على الانطولوجيا الخاصة بالصفحات، وإضافة هذه المعلومة إلى عمل تابع الاستبدال لخوارزمية الكاش المدروسة ومن ثم تقييم التطوير الحاصل على الخوارزمية، حيث يبين الشكل (1) أسماء الخوارزميات الأساسية وفق كل معيار مع إضافة المعيار الجديد المقترح وهو التشابه الدلالي، وتمثل المناطق المتداخلة في الدوائر الخوارزميات التي تعتمد أكثر من معيار لتابع الاستبدال.



الشكل (1) تصنيف خوارزميات كاش الويب

7-3-الخوارزميات المستخدمة في النظام

تبحث الورقة في تطوير خوارزمية GDFS، ولشرح الخوارزمية سوف نعطي لمحة عن خوارزمية Greedy Dual التي تعتبر الأساس الذي بنيت عليه خوارزمية GDFS.

إنّ خوارزمية Greedy Dual هي تعميم للخوارزمية المعروفة LRU (Least Recently Used) مراعيةً احتياجات التخزين المؤقت لأغراض الويب. تقوم الخوارزمية على مبدأ الاحتفاظ بقيمة تقديرية ولتكن H(p) لكل مستند أو صفحة p تم تخزينها.

عندما يتم تخزين المستند تسند لقيمته التكلفة التي حصلت نتيجة تخزين p وهي C(p)، وعندما تبرز الحاجة إلى إزالة مستند من التخزين، فسيتم اختيار المستند ذو قيمة H الأدنى، وطرد هذا المستند من الكاش وإنقاص قيمة H لكل مستند متبقي في التخزين بمقدار قيمة H للمستند المطرود. في أي وقت يتم فيه طلب مستند p تم تخزينه فإن قيمة H(p) تعاد إلى C(p). إذاً، المستندات التي تم طلبها منذ زمن بعيد تمّ إنقاص قيمها ويمكننا القول حرفياً أنه تم إعادة تصنيفها عمرياً مثل غيرها من المستندات الأخرى، وبشكل مغاير، فإنّ المستندات التي تم طلبها حديثاً يتم تهيئتها إلى قيمها الافتراضية.[6]

وجه الاختلاف بين Greedy Dual وخوارزمية LRU هو أنّ خوارزمية LRU تأخذ بعين الاعتبار فقط زمن آخر نفاذ للصفحة (طلب للصفحة)، بينما في خوارزمية Greedy Dual فالأمر مختلف حيث أنها تعتمد على قيمة المستند وعلى قيم المستندات المحذوفة سابقاً من التخزين.

إنّ GDFS هي تطوير لخوارزمية Greedy Dual السابقة من خلال إضافة تردد طلب المستند F(p) وحجم الغرض S(p) إلى قيم المفتاح Key[7].

إنّ قيمة المفتاح للمستند P تحسب وفق المعادلة التالية:

$$K(p) = L + F(p) * C(p)/S(p) \quad (1)$$

عندما يتم استعادة المستند p إلى الذاكرة لأول مرة فإن قيمة التردد الخاص به F(p) تكون مساوية لـ 1، إذا تمّ طلب النفاذ إلى p مجدداً في الذاكرة فإن قيمة التردد الخاص به تزداد بـ 1:

$$F(p) = F(p) + 1 \quad (2)$$

4-7-التشابه الدلالي Semantic Similarity

يعتبر اكتشاف الإسقاطات بين بنى المفاهيم واحداً من أصعب المشكلات التي تواجه تقنية الويب الدلالي، تصبح هذه المشكلة أصعب في المجالات التي تكون فيها المفاهيم غامضة وغير معرّفة بشكل جيد ولا يمكن إعطاؤها تعريفاً واضحاً، ولذلك تبرز الحاجة إلى إيجاد طريقة مناسبة لتمثيل المفاهيم وقياس التماثل بينها ونبين فيما يلي مجموعة من طرق دراسة التماثل بين المفاهيم الدلالية:

المقاييس المستندة على البنية أو الحواف Structure-Edge-Based Measures

النموذج المستند على الحواف يعرّف المسافة كمفهوم لقياس التشابه بين المفاهيم من خلال حساب طول المسار بين مفهومين ضمن خريطة مفاهيم (انطولوجيا) معينة. يكون المفهومين متشابهين إذا كانت المسافة بين المفهومين قصيرة وإلا يكونان غير متشابهين. يعطى التشابه المعتمد على الحواف بالمعادلة الأساسية التالية[8]:

$$\mathrm{sim}_{\mathrm{path}}(c_1, c_2) = \frac{2 * \mathrm{deep\_max} - len(c_1, c_2)}{} \quad (3)$$

- $len(c_1, c_2)$ هي طول المسار الأقصر بين المفهوم c_1 و المفهوم c_2
- $deep\_max$ هي القيمة العظمى لعمق شجرة المصطلحات ضمن خريطة المفاهيم
- $sim(c_i, c_j)$ هي التشابه الدلالي بين المصطلح c_i و المصطلح c_j

المقاييس المعتمدة على الميزات Feature-Based Measures

تلعب ميزات المصطلح دوراً مهماً في حساب التشابه الدلالي وهي تقيس المسافة بين مصطلحين بناءً على خصائصها(مثلاً تعريف المصطلح) أو العلاقة بينها ضمن وحدة تصنيف معينة (قاموس أو موسوعة) وهي تعطى بالقانون الأساسي [9]:

ليكن $\Psi(a)$ و $\Psi(b)$هي مجموعة الميزات الخاصة بالمصطلحين a و bعلى التوالي. وليكن $\Psi(a) \cap \Psi(b)$هي التقاطع بين مجموعتين الميزات و $\Psi(a) \backslash \Psi(b)$ هي المجموعة التي يتم الحصول عليها عند حذف عناصر $\Psi(b)$من مجموعة ميزات المصطلح a والتي هي $\Psi(a)$ عندها التشابه بين a و b يحسب كتابع لـ $\Psi(a) \cap \Psi(b), \Psi(a) \backslash \Psi(b)$ و $\Psi(b) \backslash \Psi(a)$على الشكل التالي:

$$sum(a,b) = \alpha . F\left(\Psi(a) \cap \Psi(b)\right) - \beta . F\left(\Psi(a) \backslash \Psi(b)\right) - \gamma . F(\Psi(b) \backslash \Psi(a)) \quad 4)$$

حيث F هو تابع يعكس أهمية مجموعة ميزات معينة وكل من α، β، γ بارامترات تحدد وزن مساهمة كل مكون.

الطرق المعتمدة على محتوى المعلوماتInformation Content-Based Measure

الطرق المعتمدة على محتوى المعلومات تستخدم محتوى المصطلح من المعلومات من أجل حساب التشابه، وهذا المحتوى يبيّن تردد وجود المصطلح ضمن نص أساسي مرجعي corpus. إن المعلومات المشتركة بين

مصطلحين تقاس بكمية المعلومات الموجودة في المصطلح المشترك الذي يشمل المصطلحين، يعطى التشابه المعتمد على المعلومات بالقانون الأساسي التالي[8]:

$$sim(c_1, c_2) = max_{c \in S(c_1,c_2)}[-\log p(c)] \quad (5)$$

حيث $S(c_1, c_2)$ هي مجموعة المفاهيم التي تتضمن كلا من c_1 و c_2

$$p(c) = \frac{freq(c)}{N} \quad 6)$$

N هو عدد الكلمات الموجودة في النص المرجعي، $freq(c)$ هو تردد وجود الكلمات المتعلقة بالمفهوم c في النص المرجعي.

الطرق المعتمدة على عدد الصفحاتPage Count-Based Measure

تعتمد هذه الطرق على عدد الصفحات التي يعيدها محرك بحث معين كنتيجة للبحث لحساب التشابه الدلالي بين الكلمات، وعند استخدام محرك البحث Google تسمى هذه المسافة بمسافة غوغل المقيسة لقياس التشابه بين الكلمات.[10]

تعرف مسافة غوغل المقيسة بالقانون:

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (7)$$

حيث f(x),f(y) هي عدد نتائج بحث غوغل للمصطلحات x,y على التوالي، f(x,y) عدد نتائج بحث غوغل عن كل من x,y معا، N إجمالي عدد الصفحات التي تم فهرستها من قبل محرك البحث.

تمّ اعتماد مسافة غوغل المقيسة لدراسة التشابه الدلالي كونه يعتبر من أحدث مقاييس التشابه ضمن تصنيفات المقاييس المذكورة أعلاه، كما أعطى نتائج أفضل لتحديد التشابه الدلالي وفق الدراسات المرجعية التي تم العودة إليها ومنها [13][12][11].

النتائج والمناقشة

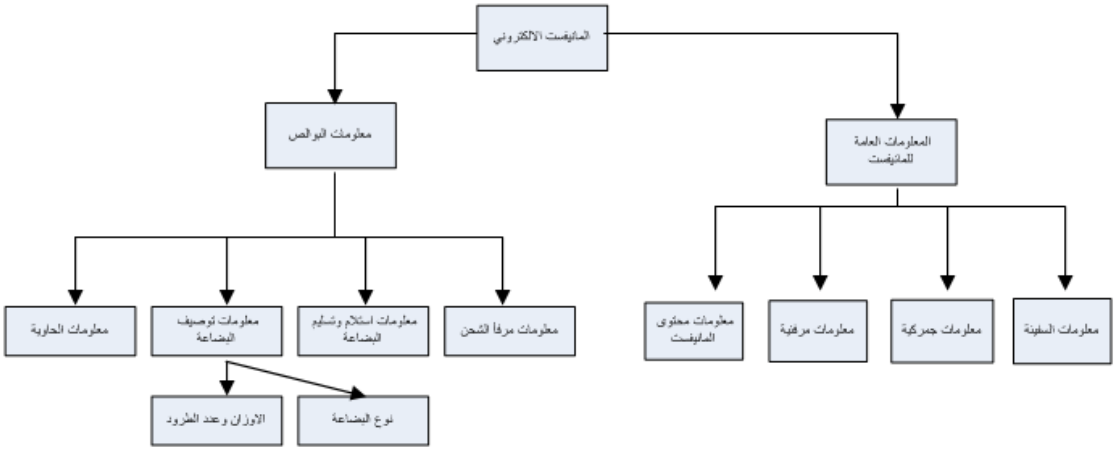
النظام المدروس

تمّ دراسة نظام المانيفست الالكتروني في الشركة العامة في مرفأ اللاذقية، حيث يعتبر المانيفست هو الوثيقة الأساسية التي تضم آلاف البيانات اللازمة لبدء العمل على تناول البضاعة من السفينة وتسليمها للزبون.

ترتبط شركة المرفأ في مديرية الجمارك في دمشق من خلال نظام إلكتروني يتم من خلاله استيراد وثيقة المانيفست من الجمارك إلى النظم المعلوماتية في المرفأ ومن ثم استخدام المعلومات الموجودة في المانيفست في باقي الأنظمة، حيث يتم قراءة المعلومات الموجودة في المانيفست وإضافة المعلومات التشغيلية عليها. يبين الشكل (3) الواجهة الرئيسية لعرض المانيفست في النظام المدروس.

الشكل (3) صفحة الويب التي تمثل الواجهة الرئيسية لنظام المانيفست في مرفأ اللاذقية

تمّ تقسيم وثيقة المانيفست إلى مجموعة الأغراض المبينة في الشكل (4)، حيث يمثل كل غرض وحدة معلومات يمكن استخدامها بشكل مستقل في الأنظمة الأخرى الموجودة، وتمّ تحديد العلاقات بين هذه الأغراض مثل علاقة IS–A التي تربط مثلاً بين مفهوم سفينة الحاويات التي هي سفينة Container Ship IS–A Ship وسفينة البضائع التي هي سفينة Cargo Ship IS–A Ship، وعلاقة Part–Of مثل علاقة الحاوية والبوليصة التي تضم مجموعة حاويات Container Is Part Of Bill Of Landing:



الشكل (4) مخطط أغراض وثيقة المانيفست الالكتروني

يضاف لاحقاً إلى الأغراض الأساسية المكونة لوثيقة المانيفست جميع الأغراض الخاصة بالعمليات التشغيلية في المرفأ، مثل معلومات الخزن و الوزن و إخراج البضاعة وغيرها العديد من العمليات، ونبين في الشكل (5) كيفية استخدام مجموعة من أغراض المانيفست الموجودة في الكاش من أجل عرض المعلومات الخاصة بنظام الخزن في المرفأ، حيث تمّ استخدام غرض معلومات البوليصة وغرض معلومات محتوى المانيفست بالإضافة لمجموعة أخرى من الأغراض.

الشكل (5) واجهة نظام الخزن في المرفأ

مخدم الكاش الخاص بالبحث:

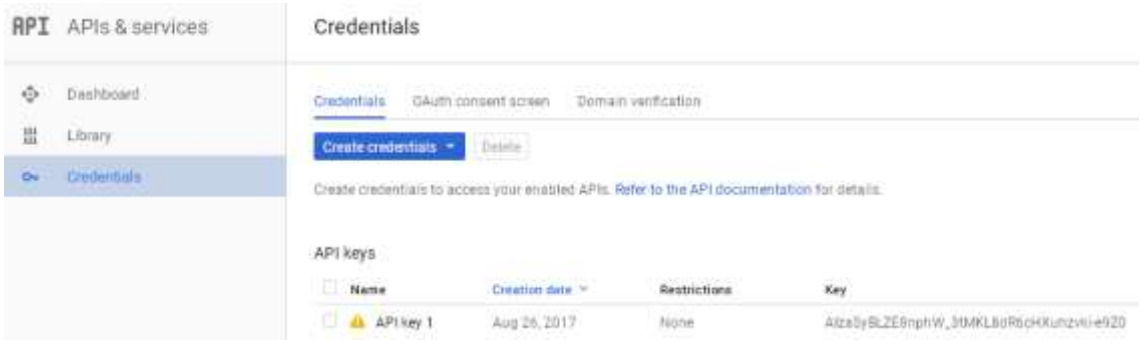
تمّ تصميم خدمة الكاش على مخدم الكاش من خلال مجموعة من الوحدات الوظيفية البرمجية التي تتصل مع بعضها من أجل تأمين وظيفة الكاش وكذلك تسجيل المعلومات التي يستخدمها البحث لتقييم الخوارزمية المدروسة. هذه الوحدات بشكل أساسي هي:

- وحدة الاتصال بقاعدة البيانات الأساسية Original Database Connection Unit: تتكون هذه الوحدة من الوحدات البرمجية التي تسمح بالاتصال مع قاعدة البيانات الأصلية بهدف قراءة البيانات من قاعدة البيانات وإعادتها إلى الزبون ومخدم الكاش.
- وحدة القراءة من الكاش Cache Reading Unit : لدى محاولة وصول الزبون إلى غرض في قاعدة البيانات الأصلية تقوم هذه الوحدة:
  - البحث عن الغرض المطلوب من قبل الزبون ضمن ذاكرة الكاش.
  - التحقق من المزامنة بين الغرض الموجود في الكاش و الغرض الأساسي في قاعدة البيانات الأساسية وفق معيار زمن حياة الغرض المحدد.
  - إعادة الغرض من ذاكرة الكاش إلى الزبون في حال وجد الغرض في الكاش.
  - استدعاء تابع القراءة من قاعدة البيانات الأصلية في طبقة الاتصال بقاعدة البيانات الأساسية، و استدعاء وحدة الكتابة إلى الكاش وذلك إذا لم يكن الغرض موجوداً في الكاش.
- وحدة الكتابة إلى الكاش Cache Reading Unitتقوم بمحاولة كتابة الغرض الذي تم جلبه من قاعدة البيانات إلى الكاش.
  - في حال وجود مكان لكتابة الغرض، يتم كتابة الغرض مباشرة إلى الكاش.
  - في حال عدم وجود مكان لتخزين الغرض المطلوب في الكاش يتم استدعاء وحدة معالجة الغرض ومن ثم تنفيذ تابع الاستبدال الخاص بالخوارزمية لتحديد الأغراض الواجب إخراجها من الكاش من أجل كتابة الغرض الجديد.

- وحدة معالجة الأغراض تقوم بعمل Word Stemming للغرض المطلوب إضافته إلى الكاش وهي العملية التي يتم بموجبها استخلاص الكلمات المفتاحية من الغرض وإزالة الكلمات المكررة والعامة مثل أحرف الجر وغيره. يتم تخزين الكلمات المفتاحية في شعاع يمثل الغرض.
- وحدة البيانات الوصفية Metadata Managing Unit: تقوم هذه الوحدة بتسجيل معلومات عن العمليات التي تمّت على ذاكرة الكاش تشمل نوع العملية (وصول – كتابة – قراءة–حذف)، وقت وتاريخ العملية بالإضافة لمعلومات الغرض الذي تمت عليه العملية، كما تقوم وحدة البيانات الفوقية بتحديث مؤشرات الغرض الذي تم الوصول إليه مثل تردد الوصول وتاريخ آخر وصول ومعلومات أخرى.

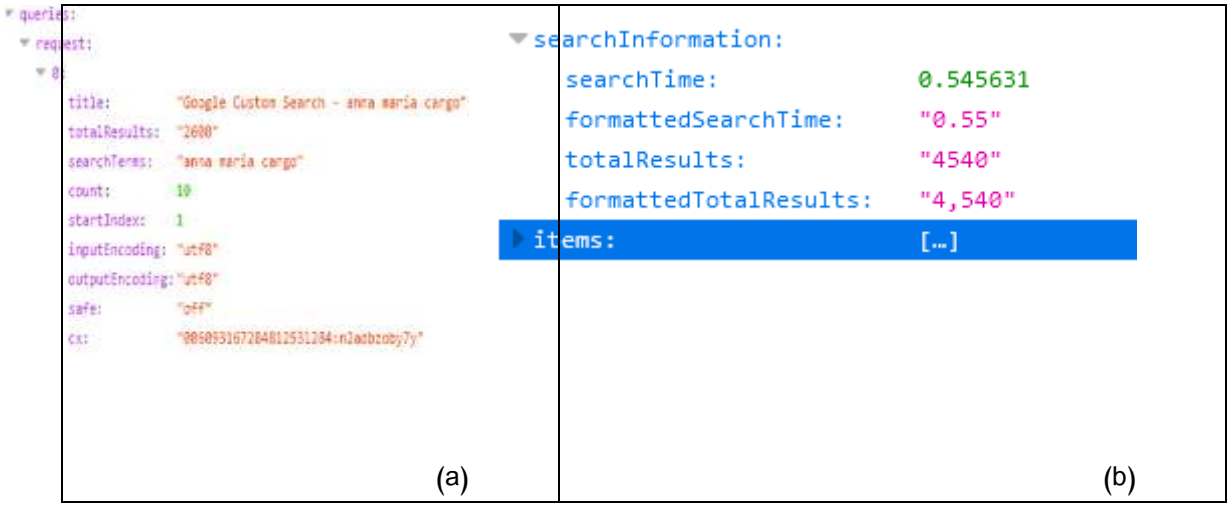
تصميم تابع الاستبدال للخوارزمية

تمّ التسجيل في موقع Google للوصول إلى خدمة Google API التي تسمح بالوصول إلى نتائج البحث عن كلمات معينة بطريقة مؤتمتة يمكن معالجتها للحصول على تردد ورود كلمة بحث معينة.



الشكل (6) التسجيل في Google API

يقوم محرك البحث Google بإعادة نتيجة البحث على شكل ملف XML كما هو مبين في الشكل (7) ويتضمن هذا الملف معلومات متنوعة مثل عدد نتائج البحث والزمن اللازم للحصول على النتيجة من محرك قواعد البيانات كما هو مبين في الشكل (7–b)، وتقوم وحدة المعالجة التي قمنا بتصميمها في تابع الاستبدال باستخلاص القيم الخاصة بتكرار الكلمة من الحقل totalresults من أجل استخدامه في حساب NGD.



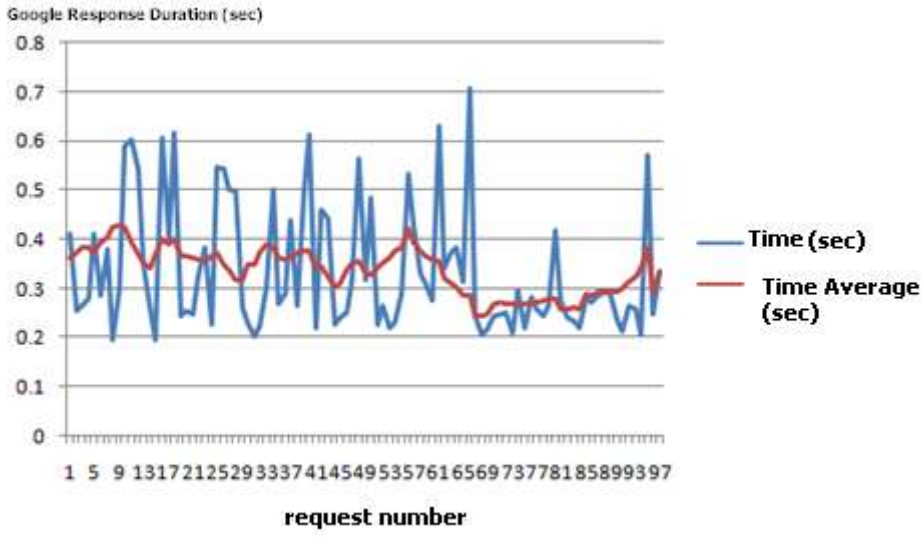
الشكل (7) مثال عن النتيجة التي يعيدها محرك البحث عن نتائج كلمة معينة

الكود التالي يبين مثال عن سلسلة البحث باستخدام محرك البحث:

https://www.googleapis.com/customsearch/v1?key=AIzaSyBLZE8nphW_3tMKL8dR6cHXuhzvkie9Z0&cx=006093167284812531284:n2adbzoby7y&q=anna%20maria%20vessel

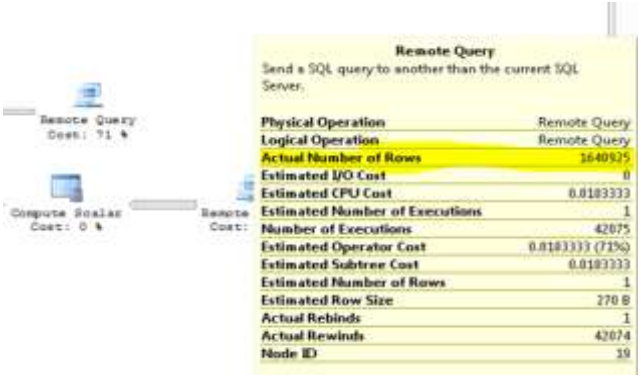
تمّ دراسة خصائص الزمن اللازم لاستعادة النتيجة من محرك البحث من خلال تطبيق البحث 1000 مرة بشكل تجريبي، وتبين لنا الخصائص التالية:

- الزمن اللازم لاستعادة النتيجة يتراوح بين 0.2 و 0.7 ثانية للبحث عن كلمتين مفتاحيتين.
- لم يتأثر الزمن بالطلبات المتوازية وعليه تمّ تصميم البنية البرمجية اللازمة لاستخدام Google API على شكل مسارات متوازية Multi-Threading لضمان سرعة استعادة النتيجة، وبالتالي، تنفيذ تابع الاستبدال بسرعة.
- متوسط الزمن اللازم لاستعادة النتيجة كما حصلنا عليه من خلال التطبيق التجريبي هو 0,33 ثانية، ويبين الشكل (8) مخطط زمن استعادة النتيجة من محرك البحث، هذا الزمن يعتبر قصير جداً نسبياً مقارنة مع متوسط الزمن اللازم لاستعادة غرض ما من قاعدة البيانات في النظام المدروس بدون استخدام الكاش وهو 13 ثانية كما هو مبين في الجدول رقم (1).
- طول السلسلة المحرفية التي يقوم محرك البحث بإعادتها لمخدم الكاش هي 75 بايت لعملية البحث الواحدة، وهو حجم صغير جداً مقارنة مع متوسط حجم الغرض في النظام المدروس وهو 253 كيلو بايت كما هو مبين في الجدول رقم (1)، ولا يشكل حركة ملحوظة على الشبكة، وهو أمر هام حيث أن أحد أهداف تطبيق الكاش هو تخفيف تبادل البيانات على الشبكة من خلال توزيع الحمل بين مخدم البيانات الأساسي ومخدم الكاش.



الشكل (8) متوسط الزمن اللازم لاستعادة نتيجة البحث عن كلمتين من محرك البحث غوغل

تمّ اعتماد كلفة الدخل والخرج I/O Cost وكلفة المعالج CPU Cost الخاصة بتنفيذ الاستعلام الذي يقوم باستعادة معلومات الغرض من قاعدة البيانات كقيمة لجزئية الكلفة في خوارزمية GDFS كما هو مبين في الشكل (9).



الشكل (9) المعلومات التي يقدمها محرك قواعد البيانات عن كلفة الاستعلام

بذلك تصبح المعادلة التي تقدم المفتاح الخاص بالأغراض الموجودة في الكاش والتي يستخدمها تابع استبدال الخوارزمية من أجل تحديد الغرض المتوجب إخراجه من الكاش لإضافة غرض جديد هي:

$$K(p) = \left(sim(o_{new}, o_j)\right) + L + F(p) * \ (\text{CPU Cost} + \text{IO Cost})/S(p) \qquad (8)$$

حيث o_{new} هو الغرض الجديد المطلوب إضافته إلى الكاش، و o_j هو غرض موجود سابقاً في الكاش و

$$sim(o_{new}, o_j) = \sum_{r=1}^{num} \sum_{q=1}^{num} 1 - NGD(k\_new_r, k\_current_q) \qquad (9)$$

حيث num هو طول الشعاع المعتمد في النظام لتمثيل الأغراض، k_new_r هي الكلمة المفتاحية ذات الترتيب r للغرض المطلوب إضافته إلى الكاش، $k\_current_q$ هي الكلمة ذات الترتيب q في الغرض الموجود حالياً في الكاش والذي تتم مقارنته مع الغرض الجديد، وبما أنّ NGD يعبر عن المسافة فإنّ $1-$NGD يعبر عن التشابه.

تكرر عملية المقارنة من أجل كل الأغراض الموجودة في الكاش لحساب الغرض الأبعد عن الغرض الحالي وذو المفتاح ذو القيمة الأدنى من أجل إزالته من الكاش.

النتائج التجريبية

تمّ اعتماد نسبة الإصابة Hit Rate وزمن تحميل الصفحة لقياس أداء الخوارزمية المدروسة، تعرّف نسبة الإصابة على أنها النسبة المئوية من الطلبات التي يمكن تلبيتها من الكاش وهي تكتب بالمعادلة التالية:

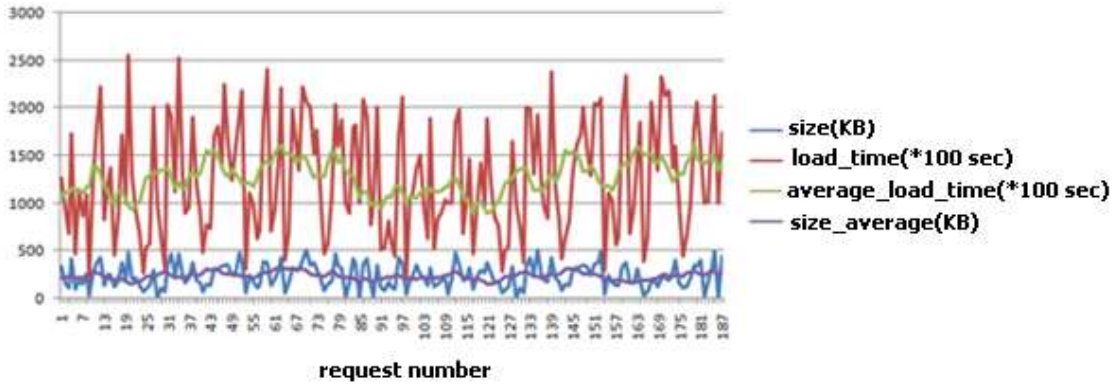
$$HR = \frac{\sum_{i=1}^{N} \delta_i}{N} \qquad (10)$$

حيث N هو العدد الاجمالي للأغراض المطلوبة و $\delta_i = 1$ اذا كان الغرض i في الكاش و $\delta_i = 0$ إذا لم يكن في الكاش.

تم حساب متوسط أزمنة التحميل من خلال الوحدة البرمجية الخاصة بالبيانات الوصفية والتي تضم عدادات زمنية تبدأ عند أول اتصال مع قاعدة البيانات في الصفحة وتتوقف عند انتهاء تحصيل المعلومات المطلوبة للصفحة من محرك قواعد البيانات، علماً أن النظم العاملة في المرفأ مبرمجة باستخدام لغة C# ومحرك قواعد البيانات SQL Server، تم تصدير النتائج الوصفية إلى برامج رسم المخططات البيانية والتي تم إدراجها في البحث.

الشكل (10) يبين عينة من مخطط أزمنة تحميل الصفحات وحجوم الصفحات الخاصة بنظام المانيفست مع المتوسطات ممثلةً بطريقة المتوسطات المتحركة البسيطة Simple Moving Average(SMA) بنافذة متحركة قدرها 10، ونلاحظ التغاير الكبير في كل من الحجم وزمن التحميل نظراً لطبيعة المانيفست، حيث يختلف طول الوثيقة حسب

اختلاف عدد الحوايا، إذ أن سفن البضاعة لا تحوي أي حاوية، أما سفن الحوايا فيتراوح عدد الحوايا بالمتوسط بين 50 إلى 1000 حاوية في المانيفست الواحد.



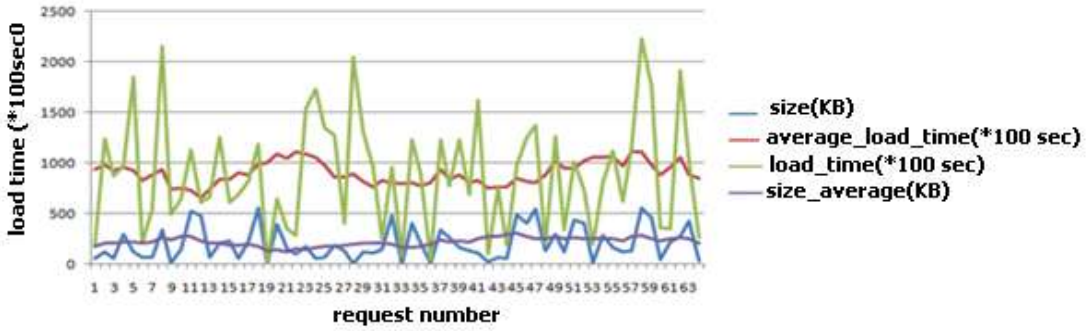
الشكل (10) أزمنة تحميل الصفحات وقياس الصفحات في النظام المدروس بدون تطبيق الكاش

الجدول (1) يبين المتوسطات للمخطط أعلاه، ونلاحظ أن متوسط زمن تحميل الصفحة هو 13 ثانية ومتوسط الحجم هو 253 كيلو بايت.

الجدول (1) خلاصة متوسطات حجوم وزمن تحميل الصفحات بدون تطبيق الكاش

	Size(kB)	Load Time(*100) sec
Average	235	1256
Max	507	2550
Min	4	184

الشكل (11) يبين عينة من أزمنة تحميل الصفحات والحجم لدى تطبيق خوارزمية GDFS بتابع الاستبدال الأساسي ونلاحظ انخفاض متوسط زمن التحميل.



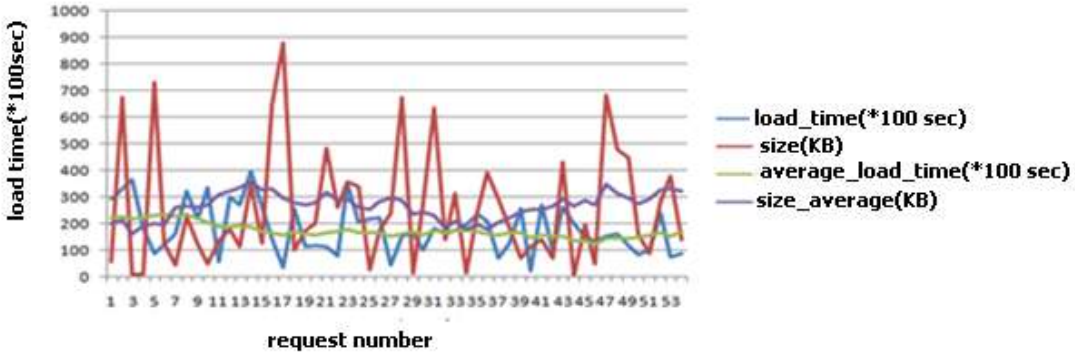
الشكل (11) أزمنة تحميل الصفحات وقياس الصفحات في النظام المدروس بعد تطبيق خوارزمية GDFS بدو تحسينات

الجدول (2) يبين المتوسطات للمخطط أعلاه، ونلاحظ أن متوسط زمن تحميل الصفحة هو 8 ثانية، ونستنتج أن مقدار التحسن في زمن تحميل الصفحة هو 40% تقريبا.

الجدول (2) خلاصة متوسطات حجوم وزمن تحميل الصفحات مع تطبيق الكاش

	Size(kB)	Load Time(*100) sec
Average	208	791
Max	558	2229
Min	6	14

الشكل (12) يبين عينة من أزمنة تحميل الصفحات والحجم لدى تطبيق خوارزمية GDFS بتابع الاستبدال المحسن بعد إدخال مسافة غوغل المقيسة، ونلاحظ الانخفاض الملحوظ في متوسط زمن التحميل.



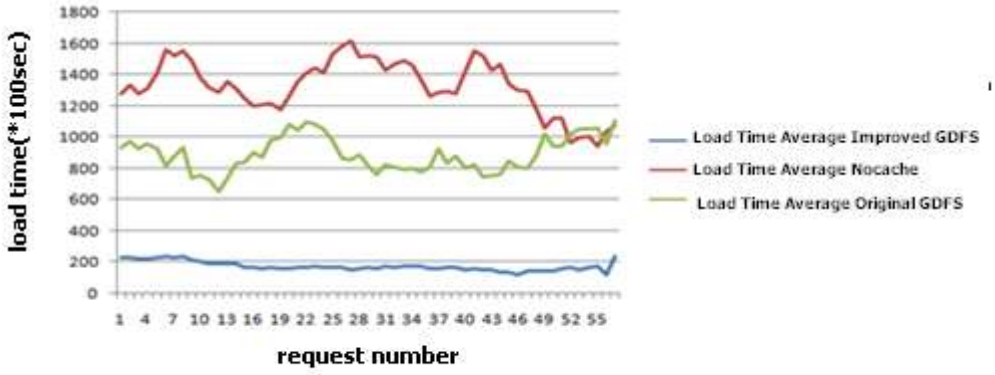
الشكل (12) أزمنة تحميل الصفحات وقياس الصفحات في النظام المدروس بعد تطبيق خوارزمية GDFS المحسنة

الجدول (3) يبين المتوسطات للمخطط أعلاه، ونلاحظ أن متوسط زمن تحميل الصفحة هو 2 ثانية، ونستنتج من خلال قيم متوسطات زمن التحميل أن مقدار التحسن في زمن تحميل الصفحة مقارنةً مع الخوارزمية الأصلية هو 75% تقريباً، ونسبة التحسن مقارنة مع النظام بدون تطبيق الكاش هو 85%.

الجدول (3) خلاصة متوسطات حجوم وزمن تحميل الصفحات مع تطبيق خوارزمية الكاش المحسنة

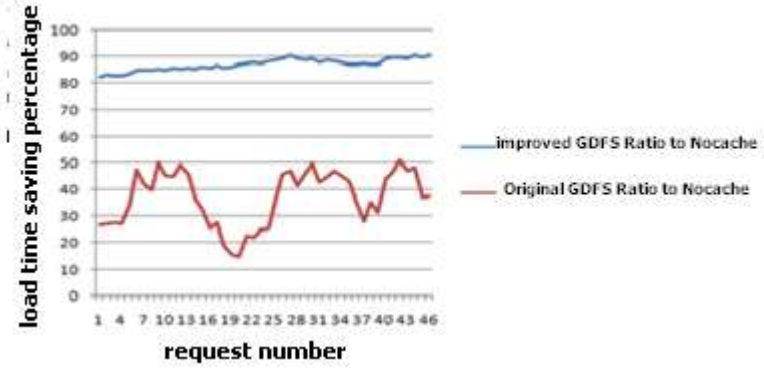
	Load Time(*100) sec	Size(kB)
Average	179	256
Max	395	880
Min	20	6

الشكل (13) يبين مقارنة بين متوسطات أزمنة التحميل.



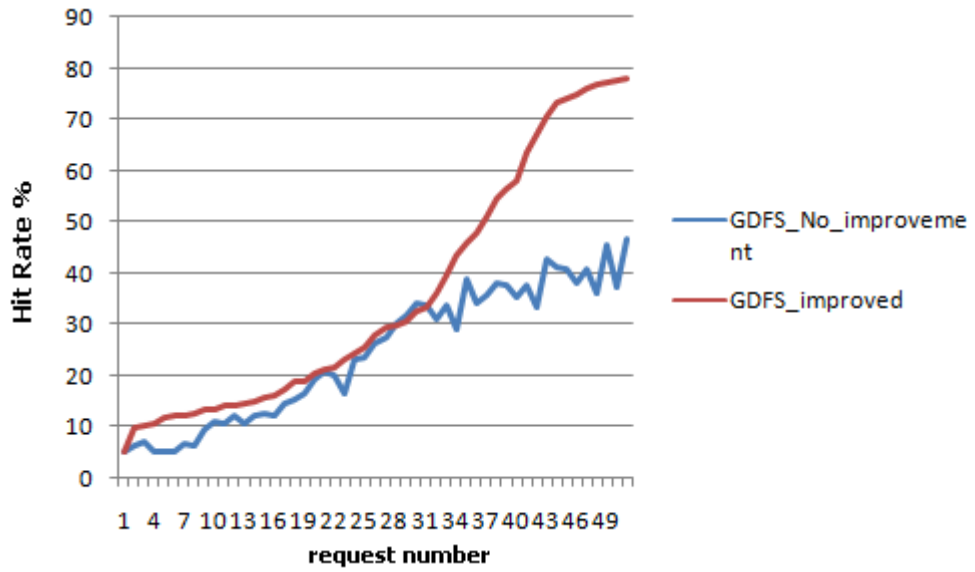
الشكل (13) مقارنة بين أزمنة تحميل الصفحات للحالات المدروسة

الشكل (14) يبين عينة من التحسين المدخل على زمن تحميل الصفحات لخوارزمية GDFS بتابع الاستبدال الأصلي، وللخوارزمية المحسنة مقارنة مع عينة من أزمنة تحميل الصفحات بدون تطبيق الكاش، ونلاحظ أن خوارزمية GDFS المحسنة قدمت تحسين كبير في زمن تحميل الصفحة وهو ما يدل على فعالية الخوارزمية في تسريع استعادة المعلومات في نظام المعلومات المدروس.



الشكل (14) مقارنة بين التحسين المدخل على زمن تحميل الصفحة بتطبيق خوارزميات الكاش

الشكل (15) يبين معدل الإصابة للخوارزمية الأصلية والمحسنة ونلاحظ بوضوح ارتفاع معدل الإصابة في الخوارزمية المحسنة، وهي تعتبر نسبة عالية مقارنةً بالدراسات المرجعية [16][15][14] التي تم العودة إليها، وهذه النسبة العالية من معدل الإصابة تفسر انخفاض زمن تحميل الصفحة نظراً لأن عمليات الحصول على الأغراض المكونة للصفحات في النظام المدروس تمت من الكاش بنسبة عالية بدلاً من قاعدة البيانات الأصلية.



الشكل (15) مقارنة بين معدل الإصابة للخوارزمية المدروسة الأصلية والخوارزمية المحسنة

الاستنتاجات والتوصيات:

- استخدام خوارزميات كاش الويب أدى إلى إدخال تحسين على زمن تحميل الصفحة بمقدار 40% بالحد الأدنى، وبالتالي ننصح بتطبيق خوارزميات كاش الويب.
- استخدام مسافة غوغل لحساب التشابه الدلالي غير مكلف زمنياً أو من ناحية حجم المعلومات التي يتم تبادلها بين موقع غوغل ومخدم الكاش الذي طبقت عليه الخوارزمية.
- إن إدخال مفهوم التشابه الدلالي إلى خوارزمية GDFS وبمقارنة متوسطات زمن تحميل الصفحات تبين حصول تحسن ملحوظ في زمن تحميل الصفحة مقارنةً مع الخوارزمية الأصلية وصل إلى نسبة 85% .
- استخدام كلفة المعالج وكلفة الدخل والخرج للحصول على الغرض مع التشابه الدلالي في تابع استبدال الخوارزمية أدى إلى إدارة ممتازة لكاش الويب بحيث تمّ الاحتفاظ بالأغراض المفيدة والمكلفة، وتبين ذلك من خلال معدل الإصابة المرتفع الذي استطاعت الخوارزمية المحسنة الوصول إليه وهو 79% والانخفاض الكبير في زمن تحميل الصفحة مقارنةً مع الخوارزمية الأصلية.

المراجع

[1] B. Moln ár, A. B,éleczki & A. Bencz úr, (2017): Information systems modeling based on graph-theoretic background,"Information systems modelling based on graph-theoretic background", Journal of Information and Telecommunication, ISSN: 2475-1839 (Print) 2475-1847 (Online),

Published online: 21 Sep 2017.

[2] Aaron Blankstein, Princeton University; Siddhartha Sen, Microsoft Research; Michael J. Freedman, Princeton University, "Hyperbolic Caching: Flexible Caching for Web Applications", This paper is included in the Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC '17). July 12–14, 2017 • Santa Clara, CA, USA,ISB N 978 -1- 931971-38 – 6

[3]الشركة العامة لمرفأ اللاذقية – مديرية المعلوماتية

[4] Minakshi Sharma , Chanda Monga, Sakshi Kalra , PG Department of Computer Science , Dev Samaj College for Women, "SEMANTIC WEB TECHNOLOGIES: AN OVERVIEW", INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

[Sharma et al.,6(1): January, 2017

[5] Muralidharan Murugesan, Dr.E.Kirubakara, Head Department of Computer Science,Nehru Memorial College,Puthanampatti Tiruchirappalli, India, Heavy Electricals Limited Tiruchirappalli, India,"Optimization of Cache Size with Cache Replacement Policy for effective System Performance", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. VI (Jul.Aug. 2017), PP 51-56

[6] Jitendra Singh Kushwah, & Dr. Sitendra Tamrakar, "An extensive Review of Webs Caching Techniques to Reduce Cache Pollution", Department of Computer Science & Engineering, AISECT University, Bhopal, India mperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-1, 2017 ISSN: 2454-1362

[7] A. Radhika Sarma and R.Go vindarajan "An Efficient Web Cache Replacement Policy",Supercomputer Education and Research Center, Indian Institute of Science, Bangalore, 560012,INDIA, In the Proc. of the 9th Intl. Symp. on High Performance Computing, (HiPC-03), Hyderabad, India, Dec. 2003.

[8] Ahmad Fayez S. Althobaiti, "Comparison of Ontology-Based Semantic- Similarity Measures in the Biomedical Text",Department of Computer and Information Sciences, Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia Journal of Computer and Communications , 2017, 5, 17- 27

[9] Nurul Aswa Omar, Shahreen Kasim, Mohd Farhan Md Fudzee, Azizul Azhar Ramli, Hairulnizam Mahdin, Seah Choon Sen," A Review on Feature based Approach in Semantic Similarity for Multiple Ontology", Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Beg Berkunci 101, 86400 Parit Raja, Batu Pahat, Johor Darul Takzim, Malaysia. Acta Informatica Malaysia 1(1) (2017) 07-09

[10] YUE FENG, EBRAHIM BAGHERI, FAEZEH ENSAN,JELENA JOVANOVIC, "A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, 2016

[11] Azadi Square, Mashhad, Razavi Khorasan,"The state of the art in semantic relatedness:a framework for comparison", Laboratory for Systems, Software and Semantics (LS), Ryerson University, Toronto, M5B 2K3 ON, Canada;Department of Computer Engineering, Ferdowsi University of Mashhad,

Department of Software Engineering, School of Business Administration, University of Belgrade, Jove, The Knowledge Engineering Review, page 1 of 30. © Cambridge University Press, 2017

[12] JUNG SONG LEE, HAN HEE HAHM ,"Less-redundant Text Summarization using Ensemble Clustering Algorithm based on GA and PSO",Division of Electronics and Information Engineering Chonbuk National University, Department of Archeology and Cultural Anthropology,Chonbuk National University,WSEAS TRANSACTIONS on COMPUTER,SE-ISSN: 2224-2872,Volume 16, 2017

[13] Panchenko, Alexander ; Fairon, Cédrick ; Morozova, Olga," A Study of Hybrid Similarity Measures for Semantic Relation Extraction". Workshop of Innovative Hybrid Approaches to the Processingof Textual Data Workshop of European Chapter of the Association for Computational Linguistics (EACL) (Avignon, France, 23/04/2012).

[14] Jie Li,Jinlong Wu, György Dán,"Performance Analysis of Local Caching Replacement Policies for Internet Video Streaming Services",Networking and Transmission Laboratory

Acreo Swedish ICT AB, Kista, Sweden,2014

[15] Sumit Rajoriya,Varsha Zokarkar,"Improving Caching Technique through Innovative Replacement Algorithm to Page for Web Proxy Caching",International Journal of Engineering Development and Research ,2016

IJEDR | Volume 4, Issue 1

[16] V.Sathiyamoorthi,"Improving the Performance of an Information Retrieval System through WEB Mining",nformation Technologies and Control, The Journal of Institute of Information and Communication Technologies of Bulgarian Academy of Sciences,Published Online: 2017-06-15 | DOI: https://doi.org/10.1515/itc-2017-0004